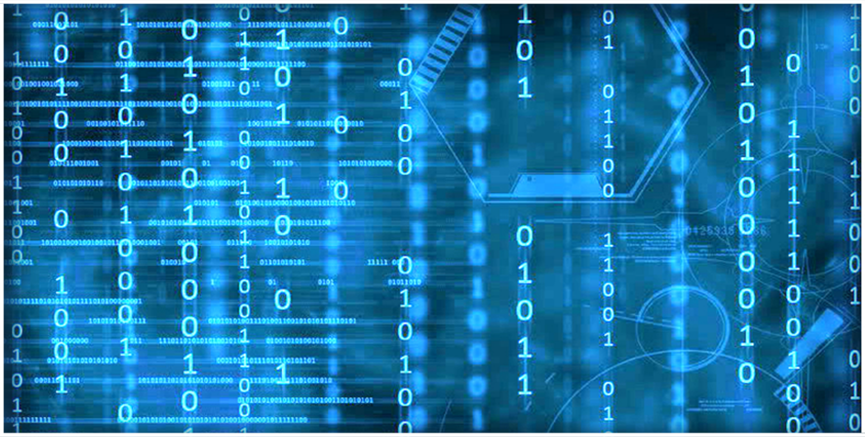
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 8 Issue 2

February 2017



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

www.ijacsa.thesai.org

SAI

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 8 Issue 2 February 2017
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Associate Editors

Reviewer Board Members

CONTENTS

Visualising Arabic Sentiments and Association Rules in Financial Text

Hamed AL-Rubaiee
Department of Computer Science
and Technology
University of Bedfordshire
Bedfordshire, United Kingdom

Renxi Qiu
Department of Computer Science
and Technology
University of Bedfordshire
Bedfordshire, United Kingdom

Dayou Li
Department of Computer Science
and Technology
University of Bedfordshire
Bedfordshire, United Kingdom

***Abstract*—Text mining methods involve various techniques, such as text categorization, summarisation, information retrieval, document clustering, topic detection, and concept extraction. In addition, because of the difficulties involved in text mining, visualisation techniques can play a paramount role in the analysis and pre-processing of textual data. This paper will present two novel frameworks for the classification and extraction of the association rules and the visualisation of financial Arabic text in order to realize both the general structure and the sentiment within an accumulated corpus. However, mining unstructured data with natural language processing (NLP) and machine learning techniques can be arduous, especially where the Arabic language is concerned, because of limited research in this area. The results show that our frameworks can readily classify Arabic tweets. Furthermore, they can handle many antecedent text association rules for the positive class and the negative class.**

Keywords—Opinion mining; Stock market; Twitter; Saudi Arabia; Association text rules; Data mining, Text Visualization

I. INTRODUCTION

The most recent research studies have been particularly interested in efforts to visualize Arabic texts. For example, Hammo et al. developed a visualization system for analysing and visualising Arabic text (VistA) [1]. Their work was based on Obeid et al.'s study [2], which applied latent semantic indexing (LSI) as a dimensionality reduction technique that aimed to stem out data from Arabic documents. By contrast, this work will use a different approach based on a hybrid of natural language processing (NLP) and machine learning algorithms involving two modules. The first module will categorize tweets as positive or negative in accordance with their sentiment polarity. The second one will help to create models useful for sentiment visualization, referring to either the positive or the negative category. Thus, this research will consider two main areas: sentiment analysis in the Arabic language and text association rules.

Similarly, Kotsiantis and Kanellopoulos [4] have demonstrated the relationships between the objects in a transactional database. Their incidence in the database defined their relationships. However, the main limitation of association rule discovery lies in its manageability when the numbers of transactions start to increase. That results in the emergence of different sets of classified data, in which most occurring objects are assigned with others in all possible ways, some of which are irrelevant.

The application of machine learning to text mining has resulted in a number of tools that are now widely implemented in different areas of research. The strongest aspect of text mining is that it is capable of processing successfully unstructured data such as social networking, e-learning, bioinformatics, pattern matching, and sentiment analysis. It also searches for and identifies patterns in data. Text mining works successfully with PDF files, emails, and XML [5].

With the spread and the rise in the popularity of social media, sentiment analysis has become one of the core social media research techniques [6-8]. Recent research studies have applied sentiment analysis to the extraction of users' views on different topics, from politics to management. The technique works well at identifying whether sentiments are positive or negative and how they are expressed [9].

In their research, Wong, Whitney, and Thomas [10] stated that the association rule in data mining was based on the inclusion of the form X→Y, where X was a set of preceding items and Y was the resultant item. It is, however, quite challenging to visualize associations in large sets of data, when over dozens of rules emerge.

This paper is organized as follows: Section 2 will discuss the sentiment analysis technique and text association rules. Section 3 will describe the methodology related to the process of the sentiment analysis of Arabic tweets and the extraction of text association rules. Section 4 will analyse the experimental findings and the visualization process. The final section will constitute the conclusion and recommendations for further work in this area.

II. RELATED WORK

A. *Sentiment Analysis in the Arabic Language*

There is scarce research on Arabic sentiment analysis, and that field is still in the initial stages of development. Arabic is very different from other languages. It has a unique structure and its own rules. For example, sentences are written from right to left, there are no capital letters, and there are a number of grammatical rules [11].

Sentiment analysis or opinion mining has been successfully applied to social media. It uses a combination of NLP and text mining to classify sentiments as positive or negative. For example, Duncan's and Zhang's research [15] specifically touched on Twitter sentiment analysis, which the nature of

Twitter and tweets reinforced. For example, spelling mistakes and the use of slang are very typical for Twitter. Also, a tweet has length restrictions; overall, it should not exceed 140 characters. Thus, the use of classification in this case was very challenging. The findings show that the level of accuracy of the neural networks in one of the experiments was relatively low. In fact, sentiment analysis applied to social media is quite different from traditional text mining. The traditional text analysis technique is based on using initially pre-defined classes to form categories in a document, and, thus, it forces the data into already existing themes [13].

The objective of sentiment analysis is entirely different. It seeks to develop new categories with regard to participants' opinions and views. The strength of sentiment analysis lies in its capacity to measure scores of sentiments by comparing that with a dictionary. Despite the uniqueness of the method, sentiment analysis has focused mainly on English text. Using the same sentiment lexicons on any other language would result in adaptation errors [12].

Machine learning (ML), which is also known as a corpus-based technique, is a supervised method in which data sets are labelled positive or negative and represented in feature vectors. These vectors, in turn, are used as training data to identify and categorize specific features in a certain class [14].

B. Text Association Rules

Chen et al. have described an association rule as an implicative insinuation of the form A ⇒ B, where A and B are frequent item sets in a transaction database and $A \cap B = \emptyset$. In practical usage, the rule is A⇒ B [16].

Text mining is withal defined as text data mining or erudition revelation from textual databases. Sodality rules are engendered by analysing data for frequent if/then patterns and utilizing the criteria support and confidence to identify the most consequential relationships. Support is a denouement of how frequently the items appear in the database. Confidence indicates the number of times the if/then representations have been found to be true. An integrated framework called associative classification was proposed that purposed to discover a set of rules that satisfied user-specified minimum support and minimum confidence as a classifier. This was done by fixating on a special subset of association rules, whose right-hand-side was restricted to the classification class attribute. The frequent if/then patterns were mined utilizing methods such as the Apriori algorithm, the Classification Based on Associations (CBA) algorithm, and the FP-Growth algorithm [8, 9].

Lopes et al. [5] expounded the quandary of mining association rules from text. They commenced by representing the text as bag of words: Let I=i_(1,) i_(2,….,),i_m. and Let D are a set of transactions, where each transaction T is a set of items that represent the document so $T \subseteq I$. An association rule is an involvement of the form $X \Rightarrow Y$ where $X \subset I$ and $Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds confidence if the document D contains X and Y, and support if the document contains $X \cup Y$. The left of the rule is the head of the rule and the set of residual words is the rule body.

Tan, Kumar, and Srivastava [7] described several key properties that should be considered to quantify the correlation between data attributes. For instance, they described the sensitivity of quantification to the row and the column scaling operation. They reported that metrics such as support, confidence, lift, correlation, and collective strength caused conflict regarding the interestingness of the pattern, and the correct metric to be used was seldom recognized.

Association rule mining can be divided into two phases. In the first phase, frequent patterns are mined with regard to the threshold minimum support. In the second phase, association rules are created with regard to the confidence threshold and minimum confidence [11].

III. METHODOLOGY

The major target of this paper is the extraction of the association rule and the visualization of positive and negative sentiments in financial Arabic text. In general, our frameworks will commence with the following:

- The Arabic sentiment analysis model,
- Pre-processing of the Arabic text,
- Classification as positive and negative sentiments,
- Model evaluation.

After the text classification, the second framework will proceed as follows:

- The creation of the Arabic text association rule model,
- The pre-processing of the Arabic text,
- Finding the frequent item set,
- Creating and visualizing the Arabic association rules for each class.

Figure 1 summarizes the first model, which was the process of opinion mining Arabic tweets. In trading strategies on the Saudi stock market, Twitter was chosen as a platform for opinion mining to illustrate the association text rules involved in Modern Standard Arabic (MSA).

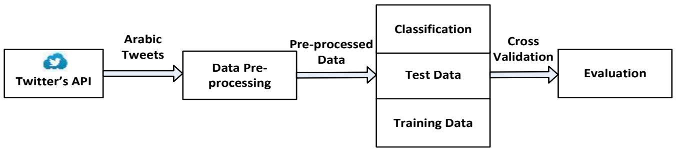


Fig. 1. The Process of Opinion Mining Arabic tweets

A. Data collection

The tweets were obtained from Mubasher firm's Twitter Account. Mubasher is high-ranking stock analysis software in the Kingdom of Saudi Arabia (KSA) and the Middle East. The tweets were gathered over a one-month period from March 1, 2016, to April 1, 2016. The data set includes 2,590 tweets, which cover most of the quota sectors of the Saudi stock market. A selection of over 100 terms and expressions in Arabic from the emotion corpus (for instance, increase, magnification, decline, fall, elevate, cash dividends, distribution of bonus shares, not to distribute) was then divided

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

between two classes: positive and negative. The most prevalent MSA words (including the following) fell into these the two classes:

(نمو Growth), (تراجع Decline), (ارتفع Rise), (انخفض fall), (ازدهار Prosperity), (تقلص Shrink), (ربح Profit), (خسارة Loss), (توزيع ارباح Dividend), and (ارتفاع خسائر High losses) .

To accumulate Arabic tweets in the corpus of data, a desktop application was developed utilizing C# and Twitter's official developers' API. The tweets utilized in the study did not involve hashtags, links, or special characters. Tweets that were duplicates or retweets were eliminated.

Three Mubasher workers who have experience with Saudi Stock Shares annotated the data manually. Negative tweets were given the label '-1', while positive tweets were given the label '1'. Neutral tweets were ignored, and the impertinent tweets were expunged from the data set.

Table 1 illustrates the number of tweets in the data set: In total, 2,590 Arabic tweets were accumulated, 934 marked tweets were utilized for the training dataset, and 1,656 extraneous tweets were erased from the data set.

TABLE I. NUMBER OF TWEETS

Positive	**Negative**	**Total**
467	467	934

B. *Data Pre-processing*

Social media channels commonly contain words with unclear meanings; opinion mining predicated on social media is still under development. This is categorically true in situations with spelling mistakes, the utilization of emoticons and other characters that express special denouement, or the utilization of English pronunciation in association with Arabic characters. Modern Standard Arabic (MSA) will be used to gratify validation requisites for this study. These kinds of tweets consist of independent, semantic-oriented Arabic lexica, which confound the research even further. To address the challenge, the ontology of an incipient keywords process model will be established to ameliorate the text mining.

Data pre-processing includes cleaning and acclimating text for classification. The pre-processing stage has several steps, for instance, online text cleaning, white space abstraction, abbreviation expansion, stemming, stop-word abstraction, negation handling, and feature selection. The final step is called filtering, while the rest are called transformations [17].

After the data labelling was completed, Rapidminer[1] was used to replace some Arabic letters that had different shapes. For example, (إ-أ-آ-لأ-لآ-لإ-ؤ-ة-ي) were replaced with (ا-لا-ه-و-ى) to remove the diacritical marks. The five pre-processing steps below were then performed using Rapidminer:

1) Tokenization: divided each tweet into multiple token-based whitespace characters;

2) Stop-word removal process: removed the Arabic stop words;

3) Light stemming: removed the suffixes and prefixes from each token;

4) Filtering token by length: abstracted worthless terms and was set to three;

5) Setting N-gram to two: N-gram was a series of n tokens from a given text [18].

Then SVM was applied with the weighting scheme TF-IDF (Term Frequency–Inverse Document Frequency) to build a classification model that could classify tweets into positive and negative classes according to their sentiment polarity. Determinately, the evaluation was carried out using the accuracy, precision, and recall methods.

C. *Classification method*

SVM's rudimentary conception involves finding a hyper-plane, which vector $\vec{\omega}$ represents, that disunites the document vectors in one class from those in other documents. SVM has been applied successfully in many opinion mining tasks. It has outperformed other machine learning techniques due to the associated advantage. For instance, the powerful in high-dimensional spaces [19].

D. *Evaluation:*

The widely known performance metrics that were utilized to evaluate the classification results were precision, recall, and accuracy [19, 20].

- Higher precision meant fewer false positives.
    - Precision =tp/(tp+fp)
- Higher recall meant fewer false negatives.
    - Recall =tp/(tp+fn)
- Accuracy involved calculating the ratio of true results (positives and negatives)
    - Accuracy=(tp+tn)/(tp+fp+fn+tn)

Figure 2 illustrates the second model, which was the process of engendering association rules in Arabic tweets. After the implementation of the previous classification model, a text association rules framework was employed to differentiate between the text rules for each class (positive, negative).

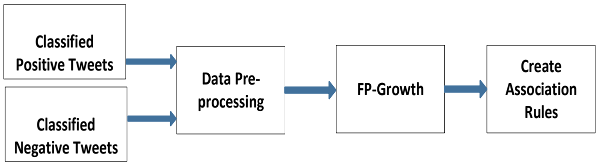


Fig. 2. The process of creating association rules for Arabic tweets

[1] https://rapidminer.com

E. Data Pre-processing

The same corpus classified as either positive or negative was used in this stage; and the same data pre-processing procedure was carried out separately for the positive class and the negative class.

F. Frequent-Pattern Method

An important algorithm in the data-mining field is a Frequent-Pattern tree algorithm. FP-growth is an approach that does not require candidate generation. It stores relevant item set information and allows an efficient novel structure to discover the frequent item sets. FP-growth has a way of decomposing the mining process into small tasks on a conditional FP-tree. First, it looks into the data set to find the frequent items at level-1 by computing the support for frequent items. Those frequent 1-item sets are stored in descending order of their supports. In the next step, the data set is scanned again to build an FP-tree using the head table with a null label root. The database scanning process continues for each transaction T to re-sort the frequent items in the header table according to the frequency of their occurrences and insert them in the FP-tree [21].

G. Create Association Rules

Association rules are if/then statements that help to expose relationships between seemingly unrelated data. An association rule has two components, an antecedent (if) and a consequent (then). An antecedent is an item set found in the data set, and a consequent is an item set found in incorporation with the antecedent [16].

IV. The Experiment Results

A. Experiment

SVM and weighting schemes (TF-IDF) were performed to explore the polarity of a given text, and to generate the word vectors. Table 2 shows the precision and recall for the SVM classifier.

TABLE II. SVM Precision and Recall

Classifier	**Accuracy**	**Recall**	**Precision**
SVM	82.31	82.31	82.75

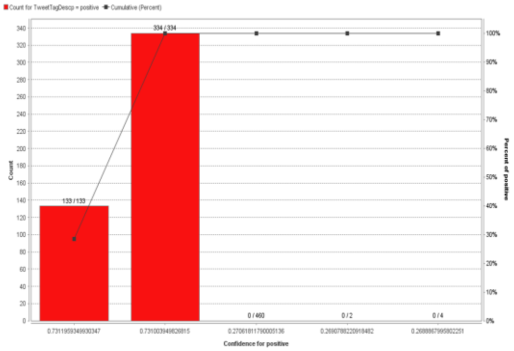


Fig. 3. SVM lift chart for the positive class

The lift chart is a way of evaluating the performance of data mining model and the predictive accuracy of one model against another [22]. The lift chart is a discrete version of representing and visualizing the classifier performance. The highest confidence numbers are shown first. As is evident, the confidence numbers decrease at some point. For example, Figure 3 and Figure 4 show the lift charts for the positive and negative classes, respectively, for an SVM classifier.

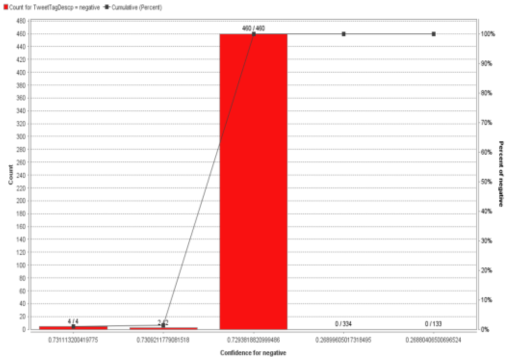


Fig. 4. SVM lift chart for the negative class

Eventually, the best precision that SVM achieved was 82.31%. On the other hand, there was virtually a 20% misclassification in our corpus. Table 3 shows the misclassification that occurred during the experiment.

TABLE III. SVM Misclassification

Accuracy: 82.31% +/- 2.93 % (mikro: 82.32%)			
	True negative	True positive	Class precision
Pred.negative	403	102	79.80%
Pred.positive	63	365	85.28%
Class recall	86.48%	78.16%	

This study aimed to continue to extract and visualize the association rules for each class because the authors believe that correlation between terms in our corpus enhanced the understanding of the text structure and clarified the sentiments expressed. This may also have improved the accuracy of our classification model.

After the separate implementation of the aforementioned association rules model for each class, the default values for most of the parameters were utilized to provide frequent items and to produce association rules that were more accurate for the positive corpus. For example, minimum support=0.01, minimum confidence=0.8, max items=2.

Experts in data mining argue that some terms or words occur with higher frequency in the dataset, while others rarely appear. In this case, the values of the minimum support will control the rule discovery. If the minimum support is set at a high value, rules that infrequently occur will not be found. Otherwise, if the minimum support is set at a low value, rules that frequently occur will be engendered. This will cause a problem called “rare item”; as a result of good rules with high confidence, it may be ignored simply because good rules have very little support [15, 23].

The frequent items produced for the positive class were 1,223. For example, Table 4 shows the frequent term that had the highest support value in the positive class with an item size equal to 1.

TABLE IV. THE FREQUENT POSITVE TERM SIZE EQUAL TO 1

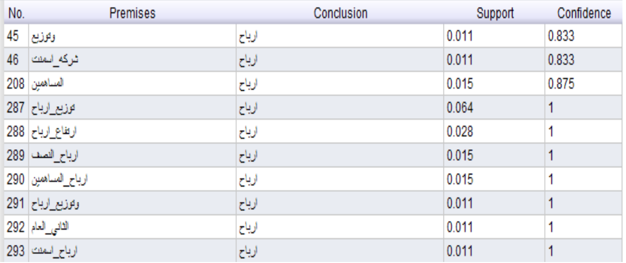
Support	Arabic Term	Term in English
0.152	الاول	First
0.131	بالربع	Quarter
0.114	تراجع	Decline
0.105	شركه	Company
0.105	تعلن	Announces

Table 5 shows some of the frequent terms or item sets that had the highest support value in the negative class with an item size equal to 2.

TABLE V. THE FREQUENT POSITIVE TERM SIZE EQUAL TO 2

Support	Arabic Term1	Term1 in English	Arabic Term2	Term2 in English
0.178	ارباح	Earnings	توزيع	Sharing out
0.155	ارباح	Earnings	تعلن	Announces
0.150	ارباح	Earnings	شركه	Company
0.124	ارباح	Earnings	زياده	Increment
0.122	ارباح	Earnings	عموميه	generality

As is evident in Figure 5, the term “earnings” (ارباح) correlated with the other terms that appeared in the premises column with the highest support values.



No.	Premises	Conclusion	Support	Confidence
45	وتوزيع	ارباح	0.011	0.833
46	شركه_اسمنت	ارباح	0.011	0.833
208	المساهمين	ارباح	0.015	0.875
287	توزيع_ارباح	ارباح	0.064	1
288	ارتفاع_ارباح	ارباح	0.028	1
289	ارباح_النصف	ارباح	0.015	1
290	ارباح_المساهمين	ارباح	0.015	1
291	وتوزيع_ارباح	ارباح	0.011	1
292	الثاني_العام	ارباح	0.011	1
293	ارباح_اسمنت	ارباح	0.011	1

Fig. 5. The term “earnings” (ارباح) and its correlations

Figure 6 shows the association between terms that were related to the positive sentiment “earnings” (ارباح). For example, the most important rules for the term “earnings” (ارباح) were [ارباح] --> [وتوزيع] (support: 0.011 confidence: 0.833), the term’s meaning entailed sharing out the profits of some company in the Saudi stock market. So the term (sharing out) was arranged together with the term “earnings” to compose positive phrases and sentences such as the following:

البنك السعودي الهولندي يوافق على زيادة رأس المال وتوزيع أرباح

Saudi Hollande Bank approves of a capital increase and dividends.

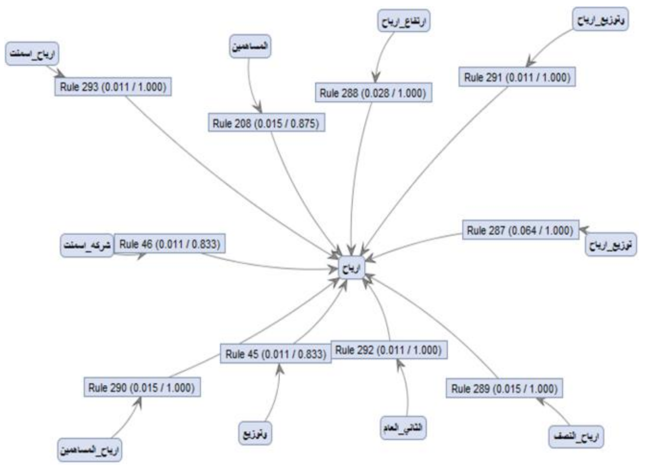


Fig. 6. Visualize the association rules term “earnings” (ارباح)

Finally, the aforementioned steps were followed to visualize the association rules for the negative class in our corpus.

The frequent items produced for the negative class were 534. For example, Table 6 shows the frequent terms that had the highest support values with an item size equal to 1.

TABLE VI. THE FREQUENT NEGATIVE TERM SIZE EQUAL TO 1

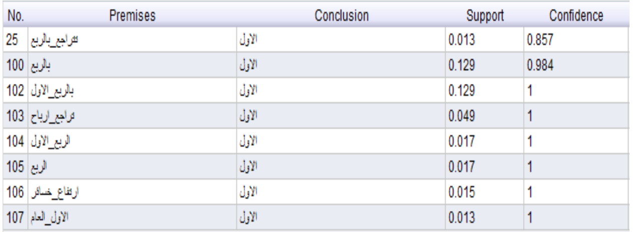
Support	Arabic Term	Term in English
0.178	ارباح	Earnings
0.155	توزيع	Sharing out
0.150	تعلن	Announces
0.124	شركه	Company
0.122	زياده	Increment

Table 7 shows some of the frequent terms or item sets that had the highest support value with an item size equal to 2.

TABLE VII. THE FREQUENT NEGATIVE TERMS SIZE EQUAL TO 2

Support	Arabic Term1	Term1 in English	Arabic Term2	Term2 in English
0.129	الاول	First	بالربع	Quarter
0.129	الاول	First	بالربع_الاول	First-Quarter
0.058	الاول	First	تراجع	Decline
0.015	الاول	First	تعلن	Announces
0.062	الاول	First	ارباح	Earnings

As is evident in Figure 7, the term “first” (الاول) correlated with the other terms that appeared in the premises column with the highest support values.



No.	Premises	Conclusion	Support	Confidence
25	تتراجع_بالربع	الاول	0.013	0.857
100	بالربع	الاول	0.129	0.984
102	بالربع_الاول	الاول	0.129	1
103	تراجع_ارباح	الاول	0.049	1
104	الربع_الاول	الاول	0.017	1
105	الربع	الاول	0.017	1
106	ارتفاع_خسائر	الاول	0.015	1
107	الاول_العام	الاول	0.013	1

Fig. 7. The correlations of the term “earnings” (ارباح)

Figure 8 shows the association between terms that were related to negative sentiments, for instance, “decline” (تراجع) and “higher losses” (ارتفاع خسائر) . For example, the most important rule for the term “first” (الاول) was [تتراجع _بالربع] --> [الاول] (support: 0.013 confidence: 0.857). The term’s meaning was dividends are decline for some company in the Saudi stock market, so the term “decline” (تراجع) was arranged together with the term “first” (الاول) to compose negative phrases and sentences such as the following:

شركه مبردتتراجع 71 % بالربع الأول من هذا العام

(Mubarrad Company *down* 71% for the first quarter of this year)

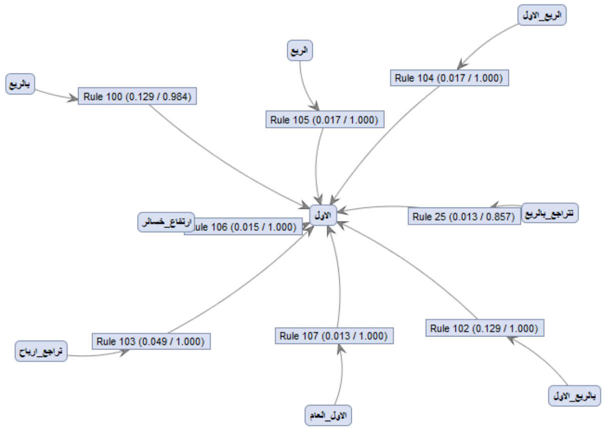


Fig. 8. Visualization of the association rules term “first” (الاول)

The experiments above were conducted on a Modern Standard Arabic (MSA) corpus and could be summarised as having used two important measures (support and confidence) to extract and visualize association rules that could expose the sentiments behind Arabic financial texts.

V. Conclusion and Future Works

In the present study, the authors designed and implemented Arabic text classifications regarding Saudi stock market opinions through the SVM algorithm and the extraction of the association rules presented. Moreover, they visualised financial Arabic text to understand the sentences’ structure and the sentiments behind them. The results of the study show that text pre-processing is an essential factor in opinion mining classification and in the extraction and visualisation of the association rules for Arabic text. Moreover, visualisation can help to sort out the misclassification that is possible with the Arabic language because of the size and ratio of the vocabulary and because of how it is characterised. In addition, as humans were involved in labelling the data, it is possible that human error occurred; for this reason, the visualisation of the text shows the importance of the correlation between terms that involved in the textual structured contents. The current study should be repeated to compare and address other metrics, for instance, lift, correlation, and collective strength. These metrics are typically used to extract and search for interesting association patterns from textual data. Several key properties should be considered in the examination of the correlations between textual data attributes in order to select the right measures for an Arabic financial domain.

Acknowledgment

This work was sponsored by Mubasher Company. We would like to thank the Managing Director of Mubasher KSA, Dr Walid Rasheed Al-Ballah, for his supplementary comments and support.

References

[1] B. Hammo, N. Obeid, and I. Huzayyen, "ViStA: a visualization system for exploring Arabic text," International Journal of Speech Technology, vol. 19, pp. 237-247, 2016.

[2] Obeid, I. Huzayyen, and B. Hammo, "Experimenting with Arabic text visualizing," in Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on, 2013, pp. 1-6.

[3] R. Elmasri and S. Navathe, "Fundamentals of Database Systems, chapter Appendix C: An Overview of the Network Data Model," ed: Addison-Wesley, 2000.

[4] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," GESTS International Transactions on Computer Science and Engineering, vol. 32, pp. 71-82, 2006.

[5] F. Gharehchopogh and Z. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in Application of Information and Communication Technologies (AICT), 2011 5th International Conference on, 2011, pp. 1-4.

[6] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in LREC, 2010, pp. 1320-1326.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Proceedings of the Workshop on Languages in Social Media, 2011, pp. 30-38.

[8] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," Expert Systems with Applications, vol. 41, pp. 3506-3513, 2014.

[9] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd international conference on Knowledge capture, 2003, pp. 70-77.

[10] P. C. Wong, P. Whitney, and J. Thomas, "Visualizing association rules for text mining," in Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on, 1999, pp. 120-123, 152.

[11] K. Ahmad and Y. Almas, "Visualising sentiments in financial texts?," in Ninth International Conference on Information Visualisation (IV'05), 2005, pp. 363-368.

[12] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, pp. 15-21, 2013.

[13] G. Fu and X. Wang, "Chinese sentence-level sentiment classification based on fuzzy sets," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 312-319.

[14] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in Collaboration Technologies and Systems (CTS), 2012 International Conference on, 2012, pp. 546-550.

[15] B. Duncan and Y. Zhang, "Neural networks for sentiment analysis on Twitter," in Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, 2015, pp. 275-278.

[16] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," IEEE Transactions on Knowledge and data Engineering, vol. 8, pp. 866-883, 1996.

[17] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science, vol. 17, pp. 26-32, 2013.

[18] G. T. a. N. S, "Feature Selection and Classification Approach for Sentiment Analysis," Machine Learning and Applications: An International Journal (MLAIJ), vol. 2, 2015.

[19] M. Rushdi - Saleh, M. T. Martín - Valdivia, L. A. Ureña - López, and J. M. Perea - Ortega, "OCA: Opinion corpus for Arabic," Journal of the American Society for Information Science and Technology, vol. 62, pp. 2045-2054, 2011.

[20] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, vol. 57, pp. 245-257, 2014.

[21] Z. Qiankun and S. S. Bhowmick, "Association Rule Mining: A Survey," Technical Repo~, CAIS, Nanyang Technological University, Singapore, p. 1, 2003.

[22] T. Jaffery and S. Liu, "Measuring Campaign Performance by Using Cumulative Gain and Lift Chart," in SAS Global Forum, 2009.

[23] B. Liu, Y. Ma, and C. K. Wong, "Improving an association rule based classifier," in European Conference on Principles of Data Mining and Knowledge Discovery, 2000, pp. 504-509.

[24] N. Obeid, I. Huzayyen, and B. Hammo, "Experimenting with Arabic text visualizing," in Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on, 2013, pp. 1-6.

[25] R. Elmasri and S. Navathe, "Fundamentals of Database Systems, chapter Appendix C: An Overview of the Network Data Model," ed: Addison-Wesley, 2000.

[26] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," GESTS International Transactions on Computer Science and Engineering, vol. 32, pp. 71-82, 2006.

[27] F. Gharehchopogh and Z. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in Application of Information and Communication Technologies (AICT), 2011 5th International Conference on, 2011, pp. 1-4.

[28] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in LREC, 2010, pp. 1320-1326.

[29] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Proceedings of the Workshop on Languages in Social Media, 2011, pp. 30-38.

[30] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," Expert Systems with Applications, vol. 41, pp. 3506-3513, 2014.

[31] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd international conference on Knowledge capture, 2003, pp. 70-77.

[32] P. C. Wong, P. Whitney, and J. Thomas, "Visualizing association rules for text mining," in Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on, 1999, pp. 120-123, 152.

[33] K. Ahmad and Y. Almas, "Visualising sentiments in financial texts?," in Ninth International Conference on Information Visualisation (IV'05), 2005, pp. 363-368.

[34] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, pp. 15-21, 2013.

[35] G. Fu and X. Wang, "Chinese sentence-level sentiment classification based on fuzzy sets," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 312-319.

[36] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in Collaboration Technologies and Systems (CTS), 2012 International Conference on, 2012, pp. 546-550.

[37] B. Duncan and Y. Zhang, "Neural networks for sentiment analysis on Twitter," in Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, 2015, pp. 275-278.

[38] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," IEEE Transactions on Knowledge and data Engineering, vol. 8, pp. 866-883, 1996.

[39] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science, vol. 17, pp. 26-32, 2013.

[40] G. T. a. N. S, "Feature Selection and Classification Approach for Sentiment Analysis," Machine Learning and Applications: An International Journal (MLAIJ), vol. 2, 2015.

[41] M. Rushdi - Saleh, M. T. Martín - Valdivia, L. A. Ureña - López, and J. M. Perea - Ortega, "OCA: Opinion corpus for Arabic," Journal of the American Society for Information Science and Technology, vol. 62, pp. 2045-2054, 2011.

[42] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, vol. 57, pp. 245-257, 2014.

[43] Z. Qiankun and S. S. Bhowmick, "Association Rule Mining: A Survey," Technical Repo~, CAIS, Nanyang Technological University, Singapore, p. 1, 2003.

[44] T. Jaffery and S. Liu, "Measuring Campaign Performance by Using Cumulative Gain and Lift Chart," in SAS Global Forum, 2009.

[45] B. Liu, Y. Ma, and C. K. Wong, "Improving an association rule based classifier," in European Conference on Principles of Data Mining and Knowledge Discovery, 2000, pp. 504-509.

Agent-based Managing for Grid Cloud System — Design and Prototypal Implementation

Osama H. Younis, Fathy E. Eassa, Fadi F. Fouz, Amin Y. Noaman
Department of Computer Science, Software Engineering and Distributed Systems Research Group, King Abdulaziz University
Jeddah, Saudi Arabia

Ayman I. Madbouly
Department of Research and Consultancy, Deanship of Admission and Registration, King Abdulaziz University
Jeddah, Saudi Arabia

Leon J. Osterweil
Department of Computer Science, University of Massachusetts, USA
Boston, USA

***Abstract*—Here, we present the design and architecture of an Agent-based Manager for Grid Cloud Systems (AMGCS) using software agents to ensure independency and scalability when the number of resources and jobs increase. AMGCS handles IaaS resources (Infrastructure-as-a-Service — compute, storage and physical resources), and schedules compute-intensive jobs for execution over available resources based on QoS criteria, with optimized task-execution and high resource-utilization, through the capabilities of grid clouds. This prototypal design and implementation has been tested and shown a proven ability to increase the reliability and performance of cloud application by distributing its tasks to more than one cloud system, hence increase the reliability of user jobs and complex tasks submitted from regular machines.**

Keywords—Cloud Computing; Grid; Management; Distributed Systems; Architecture

I. INTRODUCTION

The growing need for computational resources to solve large-scale problems leads to the cloud computing approach. Before that, Grid computing implemented a paradigm of high-throughput computing with the aim of maximal resource utilization to run multiple jobs or to solve a very big problem by parts [1]. Cloud computing now can fulfill high computations that cannot be done by supercomputers; moreover, its performance can be improved by utilizing all the available resources in a group (grid) of clouds to make sure that most resources are involved according to the required criteria. The proposed Grid Cloud computing infrastructure is considered as an emerging computing paradigm to solve complex applications in science and engineering, as it involves the combined effective utilization of cloud resources to achieve a high-performance computing, and allows the inclusion of a variety of resources like supercomputers, storage systems, and computational kernels. These resources are coupled to be available as a single integrated resource.

This infrastructure can benefit many applications, including distributed supercomputing, high-throughput computing, and data exploration. There is an increasing number of cloud providers varying in the quality of service, and the complex tasks are being increased in the fields of science and engineering. The motivation of this work is to take advantage of these clouds' services and resources to be utilized properly to execute required according to the required QoS criteria. Combining services from multiple providers give new computational capabilities, so for example instead of using costly supercomputers or HPCs, we can group Compute and Storage services (Infrastructure-as-a-Service 'IaaS') together by creating the Grid Cloud. Here, an Agent-based Manager for Grid Cloud System is presented to manage IaaS resources of grid clouds by providing an efficient way of processing high computing requests, based on software agents, for high scalability, robustness, and provider-independency. We have purchased IaaS services from two clouds to be used to execute jobs. Google Compute Engine and Windows Azure Compute have a variety of scalable services to select from, so we have integrated their APIs programmatically with our manager to be able to interact with the clouds and to execute tasks with high scalability according to the QoS. Our manager prototype has been implemented, tested and evaluated with real-world high-computing jobs

II. BACKGROUND: GRID, CLOUD AND AGENTS

Cloud computing uses remote servers and the Internet to maintain applications and data; it allows users to use applications without installation and access their data at any computer through Internet access [2], allowing for efficient computing by centralizing memory, storage, processing, and bandwidth. Cloud computing emerges from Grid computing and provides on-demand resource provisioning. A grid is usually built to utilize the idle resources in an efficient way, but the fact is that if one piece of the software on a node fails, other pieces on other nodes may fail if that component does not have a failover component on another node. Grid and Cloud computing are scalable, network bandwidth and CPU is allocated/de-allocated on demand, storage capacity is increased/decreased depending on the number of instances, users and the amount of transferred data at a given time [3-6].

The autonomous component, Software Agent, has the ability to interact with other agents and its environment on behalf of a user, capable of autonomous actions like figuring out and deciding what needs to be done to satisfy its objectives

Deanship of Scientific Research, King Abdulaziz University, grant No. (24/34 Research Group).

[7]. It is a kind of software abstraction as it provides a powerful and convenient way to describe a complex software entity, defined in terms of its behavior rather than attributes/methods in other programming languages. A multi-agent system consists of a number of interacting agents cooperating, coordinating and negotiating with one another [7]. As known, one of the cloud computing essentials is resource sharing and pooling, so in agent-based cloud computing the coordination and cooperation protocols is adopted to automate the process of resource sharing and pooling in the clouds [8].

III. Related Work

There are a number of management systems for cloud services, some of them can be found as a locally installed management application with a GUI, command-line tools, extensions of a web browser or as online tools. They provide their own management interfaces, designed to specific needs without the ability to interact with other cloud deployments of the same system, particularized to work only with a specific cloud technology and not compatible with others. Some IaaS systems are replicating the same capabilities offered by public providers like Amazon AWS; examples include Nimbus, Eucalyptus, OpenStack and OpenNebula [9, 10].

Others may only suitable for services of one cloud like [11], or for multiple services from multiple providers, like 'Karlsruhe Open Application for cLoud Administration' [12, 13]. There is also an open source, cross-platform, cloud management system called Scalr; provides server management and auto-scaling disaster recovery, where the manager is able to scale a virtual infrastructure according to the load based on RAM, disk, CPU, network or date [14]. Furthermore, there are open source initiatives like deltacloud [15], jcloud [16] and Libcloud [17], but in addition to their limitation to a specific interface, they mainly concerned with the management of public IaaS providers with basic support for private IaaS systems, while they manage virtual instances, they do not concern about the underlying physical infrastructure.

There are other related works that have involved the use of software agents in the management of the clouds, like [18]; a simulated proposed framework to manage resources for service workflows, with a hierarchical architecture for separating decisions of resource management on service, workflow and cloud levels. Another prototypal implementation found for an interface that is compliant with Open Cloud Computing Interface [19] for IaaS resources management resources negotiation, developed as an entryway to a standard FIPA multi-agent system [20]. The number of works that used software agents to manage clouds are limited. Most of them are either for resource negotiation / brokering or a simulated idea for resource allocation without implementation on real clouds [21, 22, 23]. We studied their management functions and how they were designed to understand the architectures in the computational cloud systems at the IaaS level.

The proposed concept *Grid Clouds* is similar to an existing concept called *Cloud Federation*, where services comprised from different clouds are aggregated together. In other words, the physical cloud resources are themselves being considered as a service, and cloud providers are offering their resources for other providers to expand the global cloud coverage offered to their customers without needing physical resources in every geographic locale [4]. Consequently, the cloud becomes a federation of providers/clouds that interoperate together, i.e. exchanging computing resources and data through a defined interface. There are two types of the federation, Horizontal federation and Vertical federation [24]. Horizontal federation expands the capacity of a cloud by integrating a new site and it takes place on one level of the Cloud Stack e.g., infrastructure level. Vertical federation allows the integration of new infrastructures to provide new capabilities by spanning multiple levels [24]. Presently it is almost still a theoretical concept, as there is no common standard for clouds interoperability. As an initiative for developing a common standard, the Open Cloud Computing Interface is trying to standardize an API among different clouds. This enables interoperability between providers, new business models/platforms, specialization of single clouds as well as a broader choice for users [3].

The important point here is that Cloud Federation requires one provider to rent/sell computing resources to another provider, which becomes a permanent or temporary extension of the buyer's cloud environment. Therefore, an agreement must be initiated between providers to make this federation valid. The idea of managing grid cloud services emerges new way of computing technology through grid cloud system. Related work that aggregate clouds are only simulated works; no real clouds involved in their experiments. In contrast, our manager here is using real clouds with no need for an agreement between providers, as the manager is a composite of APIs of these Clouds together to manage resources and tasks.

IV. Agent-based Manager for Grid Cloud System (AMGCS)

One way of classifying a manager is by its operation scope. A centralized manager schedules and manages all jobs submitted to the grid cloud, whereas a decentralized manager manages jobs submitted to a particular manager in the grid cloud. A centralized manager has a full knowledge and control of the resources and jobs so it can perform good scheduling, but easily become a single point of failure and a performance bottleneck. In contrast, decentralized manager architecture scales well but with low optimal scheduling performance due to the multiplicity of managers.

Scheduling policies classified into two major categories: user-oriented and system-oriented scheduling. The first is trying to optimize the performance for an individual user by minimizing the response time for each job submitted by a user, whereas system-oriented scheduling optimizes the system overall throughput and average response time [26]. A decentralized manager uses a user-oriented policy, whereas a centralized manager performs system–oriented scheduling.

The grid cloud scheduler does not own the physical resources and therefore does not have control over them [27], hence the scheduler must make best effort decision and submit the job to the resources selected. In general, the scheduler function is to map jobs to the suitable resources in the grid cloud. The scheduler involves three phases: Resources Discovery, Resource Selection and Job Execution (Fig. 1). Grid Cloud scheduling maintains a list of available resources and selects the best set of resources depending on users requirement and load balancing strategies. Then the scheduler dispatches jobs to selected virtual machines to be executed and collects the results.

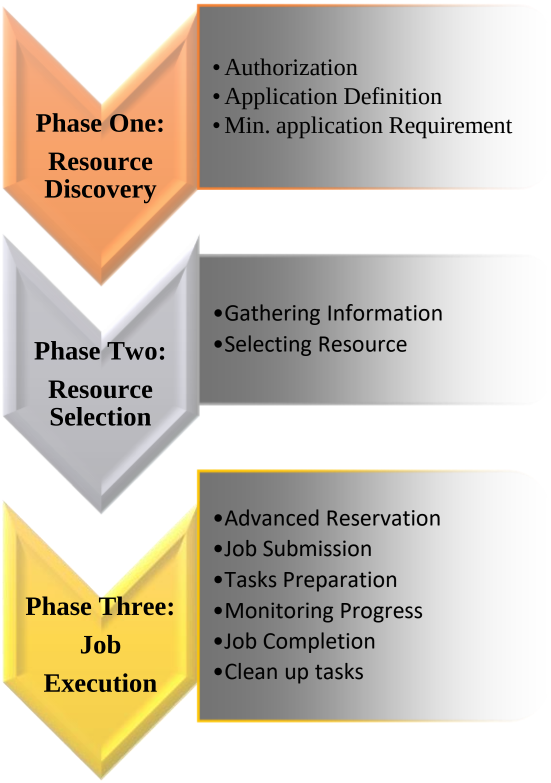


Fig. 1. Scheduling Phases

Grid Cloud resource management focuses on the virtualization and the coordinated use of heterogeneous and distributed resources. The current trend in Cloud systems is the adoption of the software agents for Grid Cloud architecture as the agents' characteristics are compatible with cloud environments [28-30]. This compatibility with Grid Cloud architecture allows us to design a manager for such architectures based on software agents to ensure platform independency and increase manager's scalability and flexibility with high cloud provider independency. Different resources in a grid cloud are varying in operating systems, CPUs, VM images, memory… etc. this difference leads to complex management for these resources. The software agent is well suited to address issues that arise from such a heterogeneous remotely controlled globally shared system. The manager here intend to group multiple clouds together (Fig. 2) and run complex jobs that need high CPU by using the CPUs of the virtual machines in the grid clouds.

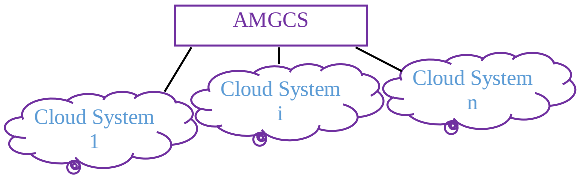


Fig. 2. Managing integrated cloud systems

The Grid cloud concept is built based on the concept of Cloud Federation, as mentioned in section 3, so a grid cloud is a way in which services characterized by interoperability features are aggregated from different clouds in one grid. It addresses the problems of vendor lock-in and provider integration, in addition to increase the performance and disaster-recovery process through techniques like co-location/geographic distribution. It also enables further reduction of costs due to partial outsourcing to more cost-efficient regions. This concept satisfies some security requirements, on un-trusted providers, by using the fragmentation technique to execute part of the job on one cloud and the other part on another cloud, then combining results without allowing each cloud to know the actual job context. Applying this concept in our manager adds benefits like resource redundancy, resource relocation and the combination of complementary services by combining different types to combined services [25]. Here we focus on the horizontal federation as it decreases provider dependency and increases availability (across multiple geographic regions). Therefore, if, for example, the QoS of executing jobs/tasks specifies a low cost, it can be executed on a cloud with the lowest cost or any other QoS requirement. Unlike Cloud Federation, AMGCS does not require an agreement between providers to integrate their services and resources, the manager itself combines the APIs of all grid clouds. The general structure of the designed Grid Cloud Manager is illustrated in Fig. 3, it consists of different agents, each of which has its own task and they are all cooperating together to achieve the manager's roles.

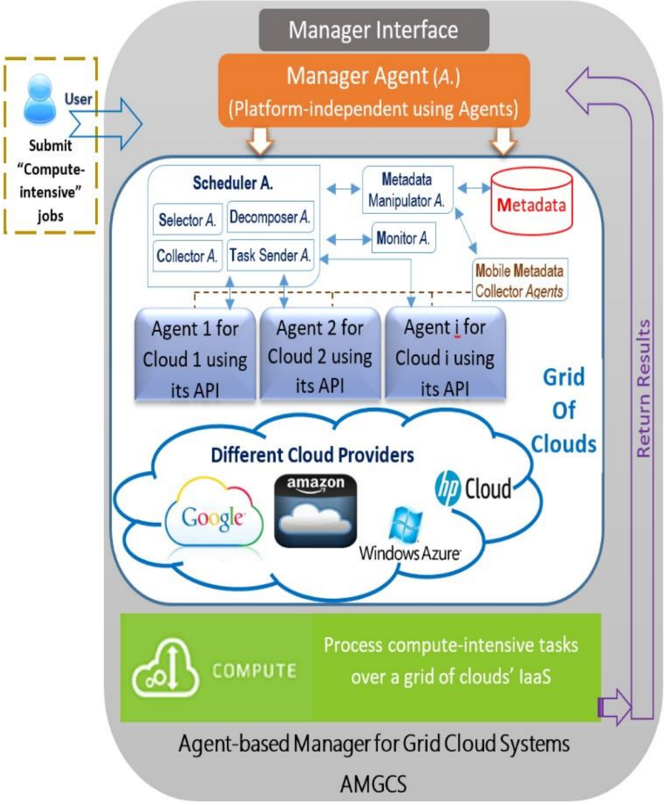

Manager Interface
Manager Agent (A.)
(Platform-independent using Agents)
User
Submit "Compute-intensive" jobs
Scheduler A.
Selector A.
Decomposer A.
Collector A.
Task Sender A.
Metadata Manipulator A.
Monitor A.
Metadata
Mobile Metadata Collector Agents
Agent 1 for Cloud 1 using its API
Agent 2 for Cloud 2 using its API
Agent i for Cloud i using its API
Grid Of Clouds
Return Results
Different Cloud Providers
Google
amazon
hp Cloud
Windows Azure
COMPUTE
Process compute-intensive tasks over a grid of clouds' IaaS
Agent-based Manager for Grid Cloud Systems
AMGCS

Fig. 3. AMGCS structure

As shown in the figure, the manager's modules are: Scheduler, Monitor, Selector, Decomposer, Collector, Metadata, Mobile metadata collector and Metadata manipulator. All of these modules are agents communicating together with a specific role for each one; the Selector is selecting resources from Metadata to assist the Scheduler in managing resources available in the grid clouds and allocating proper resources to the jobs according to the specified Quality of Service (QoS) and user desires. The monitor is monitoring jobs' executions and resources reserved on the clouds for these jobs. Task sender is the agent that invokes the API calls on the selected cloud, and hence reporting the job's and VM's status to the scheduler, as shown in Fig.4.

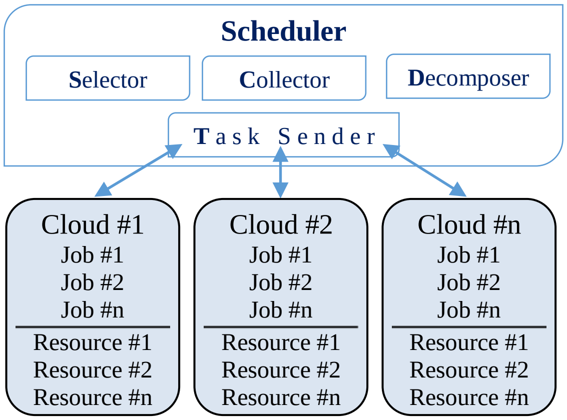


Fig. 4. Deployment of AMGCS Modules

Job decomposition helps in decomposing the job into tasks to assign each task to a proper resource according to the metadata information collected about all resources available in the grid clouds. Decomposition is done based on the job structure/nature; object-oriented system decomposed into packages, web-service based application decomposed to web services and so on. In this prototypal implementation, the decomposition is done programmatically, predefined in the application itself. The Monitor module is tracking the execution of the jobs and the associated resources, Mobile Metadata Collector agents collect and update the Metadata with available resources in the grid clouds.

AMGCS manages the virtual machines in the system and responsible for grouping multiple clouds together in the system; it schedules jobs according to the metadata information then sends the job to the selected cloud to execute it on the associated virtual machine and returns the results, finally terminates VMs after finishing their work. Therefore, the manager consists of a number of services each of which is implemented on an agent and performs a specific function of the manager's tasks, by cooperating with each other to achieve the manager's responsibilities and goals.

In order to achieve a high performance and throughput, the manager is applying best-fit and first-fit mechanisms in its selector algorithm. Best-fit shortlists the best options available for a task, this selection mechanism is slower than First-fit, which selects the first of the best. These two mechanisms are used for the purpose of ranking metadata resources and selecting the proper one that fits the task. Fig. 5 shows the algorithm used by the manager to select and manage grid cloud resources.

```
Input: Job J, User desires U, Different Resources R
Begin
Retrieve Metadata
Run Selector agent
If J is decomposable, Decompose J into tasks JTs
Create shortlist L of R from grid clouds
Sort L according to U, then apply best-fit mechanism
Create list best-fit resource BFR
Loop: select best options from L and insert into BFR
Apply first-fit mechanism on BFR
Initiate connection to the selected virtual machine VM
Task Sender agent sends J or JTs to VM
Notify Monitor agent to monitor JTs, R, VM
Execute J or JTs on selected VM
Monitor agent notifies Collector agent on completion
Collector agent returns results
Terminate connection to VM
End
```

Fig. 5. AMGCS Algorithm

V. Implementation of AMGCS

The manager has been built using Java to implement the agents that compose the manager itself. Each agent is implemented to perform specific tasks, managing resources on different clouds and updating the metadata that contains the information about the available resources in all grid clouds. The manager is integrated with multiple clouds APIs, these clouds are Google Compute Engine, Google Cloud Storage, and Windows Azure Compute. Resources on these clouds are managed/controlled through API functions of each cloud. Authorization and authentication processes must be initiated once before the actual invoking of functions and making requests.

A. Integration with Google Compute Engine (GCE)

The API of GCE has been integrated with our manager to directly call requests to manage available resources. GCE first needs to authenticate the machine before accepting any request; this is done through the OAuth 2.0 protocol. This authorization framework provides clients or third-party applications a method to access resources (HTTP services) either by allowing them to obtain access on its own behalf or on behalf of a resource owner by coordinating an approval interaction between the resource owner and the HTTP service.

Fig. 6 shows the architecture of GCE and the different ways of accessing the cloud, Command Line Interface (CLI), User Interface (UI) and API code library.

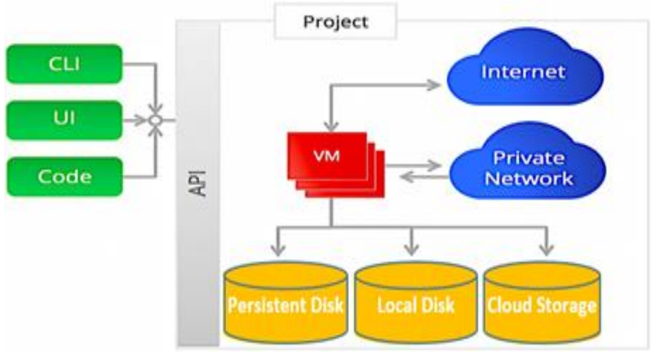


Fig. 6. Google Compute Engine Architecture

To use this API within our manager, a key is required from Google Compute Engine that must be associated with the manager in order to be authorized to perform any requests, this key is called "client_secrets" and can be downloaded in a JSON format from Google Developers Console, example client_secrets.json file:

```
{
 "installed": {
 "client_id": "837647042410-75ifg...usercontent.com",
 "client_secret":"asdlkfjaskd",
 "redirect_uris": ["http://localhost", "urn:ietf:wg:oauth:2.0:oob"],
 "auth_uri": "https://accounts.google.com/o/oauth2/auth",
 "token_uri": "https://accounts.google.com/o/oauth2/token"
  }
}
```

The manager will use OAuth 2.0 in order to authenticate the RESTful API. This API will be used to create and delete disks, virtual machine instances, and other resources, and to integrate with other Google Cloud services, i.e. Google Cloud Storage which is used here to store the metadata for the manager itself. Fig. 7 shows the flow of authenticating APIs calls.

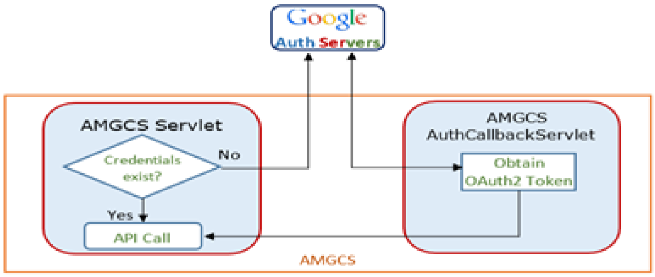


Fig. 7. Authenticated API calls sample flow

Servlet is a Java class to extend the server capabilities to respond to any requests types and to extend the applications hosted by web servers. Several machine types are available from GCE; micro, standard, high CPU and high memory machine types, shown in Table 1. AMGCS select high CPU types for tasks require more virtual cores relative to memory. GCE uses GCEU (Google Compute Engine Unit) as a unit of CPU capacity describing compute power. Minimum power of one logical core on the Sandy Bridge platform is 2.75 GCEUs.

TABLE I. SELECTED LIST OF MACHINE TYPES ON GCE

Configuration	Virtual Cores	Memory
Micro - Small	Shared core	0.60 – 1.7 GB
Standard	1	3.75 GB
	2	7.50 – 13 GB
High Memory,	4	15 – 26 GB
High CPU	8	30 – 52 GB
	16	60 – 104 GB

These machine types, in addition to many others, are included in our metadata database, so the scheduler will choose the proper one to execute the job, according to the required QoS and whether cost or response time is the most critical factor for user desires. After selecting the machine type, API request will be sent, after being authenticated, to initiate a new instance with the specified properties/configurations. Afterward, the manager calls an API function to start running the specified instance (VM) on that Infrastructure using the instances().insert function. These instances can run Linux server from many available images; provided by Google or customized images of other systems, as needed. Finally, jobs will be executed on this instance and others on other instances, results returned to the manager then to the user. The integration of GCE API with our manager is illustrated in Fig. 8.

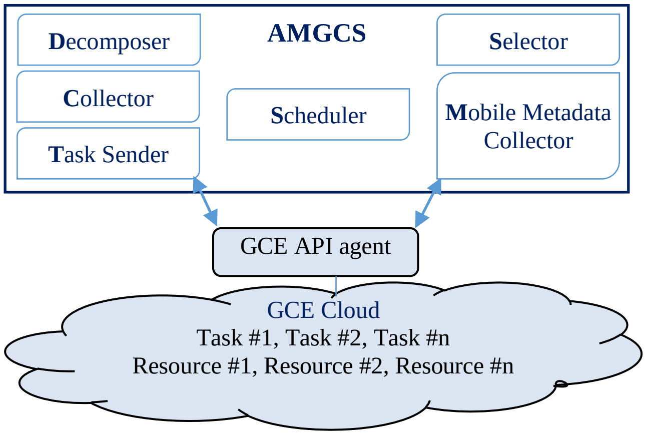

AMGCS
Decomposer
Selector
Collector
Scheduler
Mobile Metadata Collector
Task Sender
GCE API agent
GCE Cloud
Task #1, Task #2, Task #n
Resource #1, Resource #2, Resource #n

Fig. 8. GCE integration with AMGCS

B. Integration with Windows Azure Compute (WAC)

Windows Azure is supporting Microsoft operating systems and non-Microsoft operating systems. Its VM image gallery includes latest releases of Windows Server, SharePoint, SQL Server, BizTalk Server, and many non-Microsoft workloads like Ubuntu, SUSE Linux, openSUSE, OpenLogic, etc. Integrating WAC with AMGCS gives the power of handling yet more requests for high computing by calling the associated API call to create new instances, with built-in capability of Load Balancing, monitoring and restarting VMs.

Windows Azure Compute has many machine types ranging from extra small to extra-large machines, also AMGCS will select from this wide range of machine types the proper machine type with proper configurations that suit the job requirements, Table 2 shows some of these machine types, and Fig.9 shows the integration of GCE API with manager’s agents.

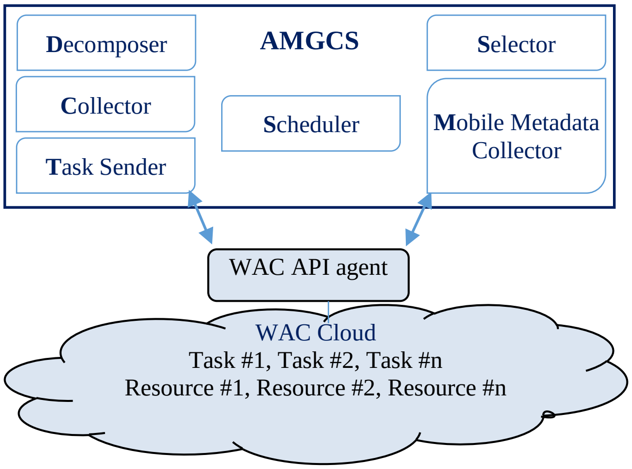

Decomposer
AMGCS
Selector
Collector
Scheduler
Mobile Metadata Collector
Task Sender
WAC API agent
WAC Cloud
Task #1, Task #2, Task #n
Resource #1, Resource #2, Resource #n

Fig. 9. Windows Azure Compute integration with AMGCS

TABLE II. SELECTED LIST OF MACHINE TYPES ON AZURE COMPUTE

Configuration	Virtual Cores	Memory
Extra small	Shared core	768 MB
Small	1	1.75 GB
Medium	2	3.5 – 14 GB
Large	4	7 – 28 GB
Extra large	8	14 – 56 GB

In order to use WAC API, the following is required:

1) A subscription Id: which uniquely identifies our subscription; this id is obtained from the Windows Azure portal.

2) A management Certificate: that must be associated with our subscription to authenticate the API calls.

As the endpoints are accessible over HTTP, we have to create, in our code, an endpoint specific to a particular kind of operation we want to perform, then create HTTP request for that endpoint and the management certificate is attached with that request to be authenticated.

C. AMGCS Metadata

The manager schedules tasks according to this updated metadata information. A cloud database is used here to store the manager's metadata; this cloud database is Google Cloud Storage, to guarantee the compatibility and independency of any platform. Provided by Google, with features like object versioning, parallel uploads and CRC-based integrity checking to maintain the robustness of our sophisticated manager. We can access its API using XML, JSON or using the libraries for several popular programming languages including Java. This storage service is used here to guarantee platform independency and the proper integration with mobile metadata collector agents. The metadata includes the following info of each cloud system in the grid clouds:

Name: Name of the resource or virtual machine.

Description: Description about the resource.

ID: The unique ID of the resource.

CPUs: Number of CPUs in the virtual machine

ImageSpace: The size of the server image in Gigabyte.

Kind: The category of the virtual machine, e.g. high memory, high CPU, or standard.

Disks: The maximum number of disks can be associated to a specific virtual machine.

DisksSize: The size of the disks associated to a specific virtual machine.

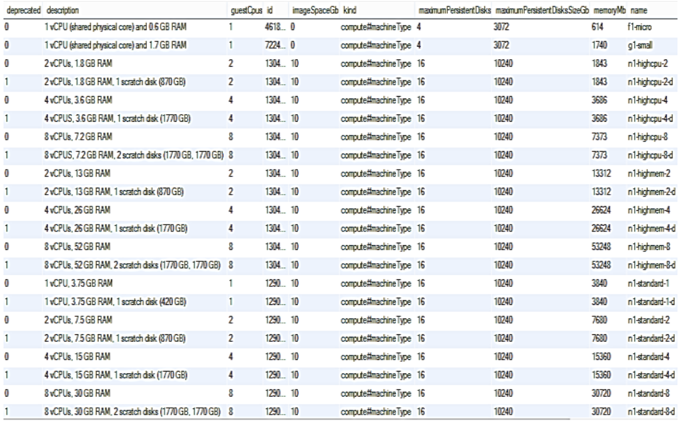
Memory: The size of memory in Megabyte.

Location: The location of the server, e.g. Central US, West Europe, East Asia… etc.

ServerType: The type of the server, e.g. Windows, Linux, SQL, Oracle…etc.

ServerImage: The image of the server, which contains the boot loader, an operating system and a root file system that is necessary for starting an instance, e.g. debian-7, centos-6, rhel-6, sles-11, Windows Server 2012 R2 Data-center, SQL Server 2012 SP1 Enterprise, OpenSUSE 13.1, Ubuntu Server 14.04 LTS… etc.

Fig. 10 is a snapshot of the current metadata collected about available resources



deprecated	description	guestCpus	id	imageSpaceGb	kind	maximumPersistentDisks	maximumPersistentDisksSizeGb	memoryMb	name
0	1 vCPU (shared physical core) and 0.6 GB RAM	1	4618...	0	compute#machineType	4	3072	614	f1-micro
0	1 vCPU (shared physical core) and 1.7 GB RAM	1	7224...	0	compute#machineType	4	3072	1740	g1-small
0	2 vCPUs, 1.8 GB RAM	2	1304...	10	compute#machineType	16	10240	1843	n1-highcpu-2
1	2 vCPUs, 1.8 GB RAM, 1 scratch disk (870 GB)	2	1304...	10	compute#machineType	16	10240	1843	n1-highcpu-2-d
0	4 vCPUs, 3.6 GB RAM	4	1304...	10	compute#machineType	16	10240	3686	n1-highcpu-4
1	4 vCPUS, 3.6 GB RAM, 1 scratch disk (1770 GB)	4	1304...	10	compute#machineType	16	10240	3686	n1-highcpu-4-d
0	8 vCPUs, 7.2 GB RAM	8	1304...	10	compute#machineType	16	10240	7373	n1-highcpu-8
1	8 vCPUS, 7.2 GB RAM, 2 scratch disks (1770 GB, 1770 GB)	8	1304...	10	compute#machineType	16	10240	7373	n1-highcpu-8-d
0	2 vCPUs, 13 GB RAM	2	1304...	10	compute#machineType	16	10240	13312	n1-highmem-2
1	2 vCPUs, 13 GB RAM, 1 scratch disk (870 GB)	2	1304...	10	compute#machineType	16	10240	13312	n1-highmem-2-d
0	4 vCPUs, 26 GB RAM	4	1304...	10	compute#machineType	16	10240	26624	n1-highmem-4
1	4 vCPUs, 26 GB RAM, 1 scratch disk (1770 GB)	4	1304...	10	compute#machineType	16	10240	26624	n1-highmem-4-d
0	8 vCPUs, 52 GB RAM	8	1304...	10	compute#machineType	16	10240	53248	n1-highmem-8
1	8 vCPUs, 52 GB RAM, 2 scratch disks (1770 GB, 1770 GB)	8	1304...	10	compute#machineType	16	10240	53248	n1-highmem-8-d
0	1 vCPU, 3.75 GB RAM	1	1290...	10	compute#machineType	16	10240	3840	n1-standard-1
1	1 vCPU, 3.75 GB RAM, 1 scratch disk (420 GB)	1	1290...	10	compute#machineType	16	10240	3840	n1-standard-1-d
0	2 vCPUs, 7.5 GB RAM	2	1290...	10	compute#machineType	16	10240	7680	n1-standard-2
1	2 vCPUs, 7.5 GB RAM, 1 scratch disk (870 GB)	2	1290...	10	compute#machineType	16	10240	7680	n1-standard-2-d
0	4 vCPUs, 15 GB RAM	4	1290...	10	compute#machineType	16	10240	15360	n1-standard-4
1	4 vCPUs, 15 GB RAM, 1 scratch disk (1770 GB)	4	1290...	10	compute#machineType	16	10240	15360	n1-standard-4-d
0	8 vCPUs, 30 GB RAM	8	1290...	10	compute#machineType	16	10240	30720	n1-standard-8
1	8 vCPUs, 30 GB RAM, 2 scratch disks (1770 GB, 1770 GB)	8	1290...	10	compute#machineType	16	10240	30720	n1-standard-8-d

Fig. 10. Manager's metadata

After the job is received by the manager, it will look for the suitable resource available from this metadata, then start a new virtual machine with specific properties on the specific cloud provider, then it will send the task to this particular virtual machine. Another agent of the manager will monitor the execution of these tasks on these resources, and will send periodic notifications to inform when the task or job execution is completed. To minimize costly communication of large amounts of data, some intercloud mapping feature is needed, so most clouds provides a feature called VHD (Virtual Hard Disk), which is a file format used as the hard disk of a virtual machine that may contain any amount of data. It is portable to be attached to any cloud virtual machine from any provider, so the manager can attach such a VHD to any created VM. After finishing the job execution, the manager will make a request to terminate the VM and deals locate the associated resources. The integration of the GCS API with AMGCS agents is illustrated in Fig. 11.

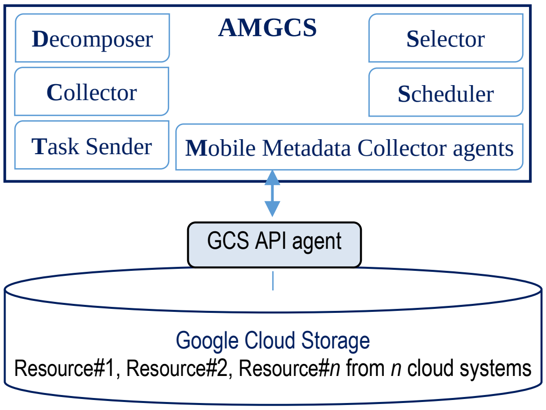

AMGCS
Decomposer
Selector
Collector
Scheduler
Task Sender
Mobile Metadata Collector agents
GCS API agent
Google Cloud Storage
Resource#1, Resource#2, Resource#n from n cloud systems

Fig. 11. Google Cloud Storage integration with AMGCS

VI. Evaluation and Testing

To evaluate AMGCS system, various testbeds have been set up to measure its efficiency, and we have successfully tested the execution of complex jobs that require high computations. If a compute-intensive job is requested to be executed with specified user desires like performance, cost or reliability, the AMGCS manager selects the proper resource from the metadata and sends the job to that resource in order to execute it. This process requires having full information about the resources available in the grid clouds, which is done through our manager by calling the API functions associated with each cloud. There is an updated list of this information in the manager's metadata.

A. *Test cases and Results*

With a centralized manager, the AMGCS has been tested with a compute-intensive job, a big matrices multiplication job. These matrices are huge and need a long time to be manipulated. The size of the first matrix is [4096][2048] and the second matrix size is [2048][2048], this calculation would take a long time to be done, depending on the type of the machine that executes the job, memory, location, server type and so on. The first test case is a single job executed on a single cloud system, with the required user desires: Low cost and minimum execution time. If a regular user wants to perform this job on a cloud system, he will just select any cloud with any properties as he is looking for a low cost, he might manually select the lowest-cost resource to execute this job. In contrast, the AMGCS manager will select the most proper resource that suits the user desires, and achieves this job efficiently.

Consider, as a case, the user has sent the job arbitrary to a lowest-cost resource, which is a virtual machine with a shared core, the job of multiplying these huge matrices took approximately 43.8 minutes to be done. However, AMGCS manager submitted this job to a more proper VM from the list of resources available in the manager's metadata, it is also a shared core but with capabilities that make this VM the best option to select from the available resources in the grid cloud, "Memory and *ServerImage*". Fig. 12 shows the result of this test case.

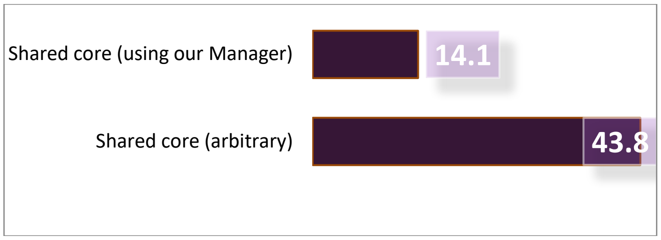


Fig. 12. Comparison of execution time (minutes), shared core

The optimized execution is shown in Fig. 12 is achieved by executing the job on a VM with proper properties (capabilities), and because the manager knows the full details about all resources from the metadata, it chose one cloud of the grid clouds and create a proper virtual machine to execute this job. The configuration of this particular VM customized the memory, and the server image (Debian 7 Wheezy). Here, obviously, one server is better than the other and hence the significant difference in execution time between them. The server type at the bottom is a Windows server, and the type of the upper (optimized) one is a Linux server.

To prove that the enhancement here is achieved by the manager's selection strategy and not by the server type (Linux or Windows), we did a second test to arbitrary execute the same job on a Linux server also with a shared core, but without using our manager. The results proved that the AMGCS scheduling is the reason for that significant enhancement, because of the many options that can be customized to a particular VM to make it the best proper option to execute the required tasks, unlike the regular user's selection that may ignore any consideration to the capabilities or properties of the server's VM. Fig. 13 shows the execution time on the other server type (Linux) without using AMGCS, it is almost near the time taken on the Windows server in Fig. 12, only four minutes less.

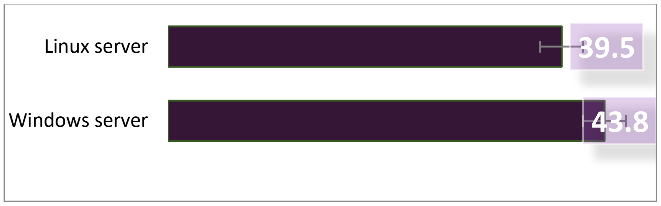


Fig. 13. Execution time (minutes) without AMGCS, shared core

A third test case, the same job but different user desire, which is minimum execution time. Here the user does not care about the cost and the most care about the execution time. It also tested by submitting this job arbitrary to any cloud with any properties and compare this with the selection of our manager which depends on knowledge of the user desires, job structure and nature, resources available and the recommended resources for complex jobs. As we need here the minimum execution time, a high CPU power is required to solve this job as fast as possible. Hence, a virtual machine with eight cores is the proper one. Yet, even with eight cores, the performance could be optimized further by taking into account factors that might degrade the performance or efficiency of execution, like memory and server type. So here the job has been executed on multiple VMs that have the same core number, eight cores. Fig. 14 shows the difference in execution time; the upper one is much faster compared to the arbitrary selected VM at the bottom. This is because the manager has submitted the job to a more suited cloud with better virtual machine capabilities (*serverType* and *memory* space).

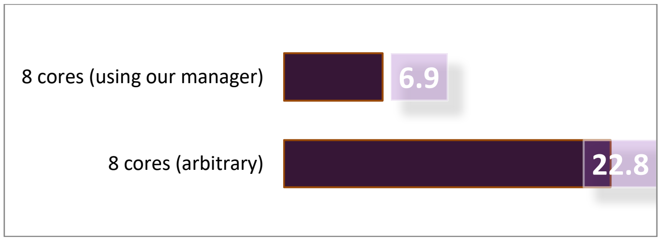


Fig. 14. Comparison of execution time (minutes), 8 cores

All these aforementioned test cases have submitted the job to the clouds' VMs without decomposing it into tasks. Decomposing this big job into tasks to be individually executed on multiple clouds will increase the performance and reliability of the job execution.

The fourth test case is to measure the improvement of enhancement after decomposing a job, the same job has been divided into two tasks (parts) to be executed on the grid cloud system, with one user desire which is minimum execution time. Decomposition here is programmed in the code just to test the prototypal manager, by dividing the first matrix by half and keep the second as it is, to maintain the matrix multiplication rules. Now the manager has many options to execute these tasks; one of these options is to send each of these tasks to a different virtual machine in order to be executed separately and then combine the results together. This is the case here, where the two tasks of the multiplication job are processed on multiple clouds from the grid clouds; hence, the time decreased by half. Fig. 15 shows the significant difference in time compared to the execution on single cloud system.

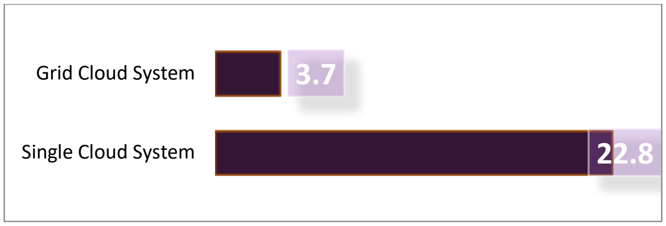


Fig. 15. Comparison of execution time 'minutes', Single vs. Grid Cloud

The fifth test case was conducted to evaluate the improvement of using grid clouds in executing tasks, by sending replications of these tasks (parts of the job) to multiple clouds. Each task has been replicated and processed two times on multiple virtual machines (other than previously used VMs) so if any failure occurs in any VM we still have another copy on another VM. Hence, the reliability of execution is guaranteed for these tasks, despite the cost that might be high because here reliability is the user desire and reliability is always costly. Fig. 16 shows the results of this experiment.

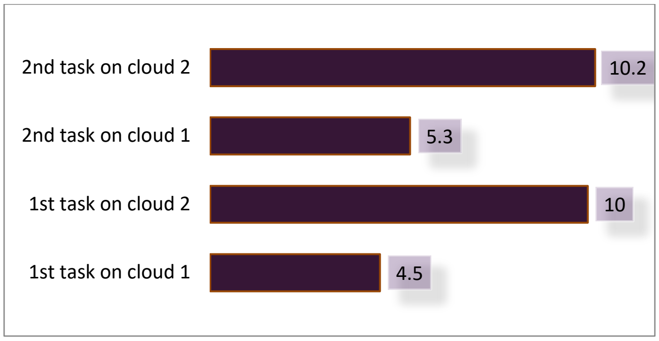


Fig. 16. Execution time (minutes) for tasks of the job, on Grid Cloud

We end up with an enhancement in executing complex tasks on grid cloud resources in an efficiently managed way. Combining comparisons above proves that there is an improvement by 16% - 30% between single and grid cloud system, illustrated in Fig. 17.

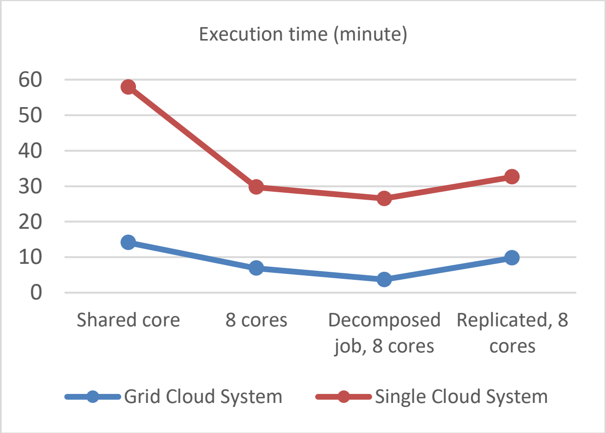


Fig. 17. Overall enhancement, AMGCS vs. Single Cloud

B. Discussion on Experiments Results

Depending on the results that we got from the performed tests and experiments, our finding could be summarized as follows:

- The Manager solves current challenges of executing tasks on the cloud, utilizes grid clouds' resources, solves complex and compute-intensive tasks, and tasks that require high reliability and high performance.
- AMGCS manages jobs and resources and gives good performance in terms of execution time, resource utilization and system throughput compared to a single cloud system.
- Increasing the number of the grid clouds in the system gives more optimize options and high performance compared to using a small number of grid clouds.
- The overall enhancement of Grid Cloud System is about 16% - 32% compared to a single Cloud System.
- The manager does not require any provider-side agreement, only configuring the libraries of the grid clouds.

- Fault tolerance is guaranteed by replication and increased performance through scaling resources to accommodate user's needs, more or less.
- There is a trade-off between high reliability and cost, our manager may replicate tasks on multiple clouds and hence more cost.

VII. Conclusion

In this paper, we introduced an agent-based manager for grid cloud system that has been designed based on software agents to ensure platform independency, heterogeneity handling and flexibility of managing grid clouds. It has been designed, implemented and successfully tested on real clouds. The limitations of the proposed manager could be summarized in its inability to be fully interoperable between different virtualization technologies and recourses compatibilities from different providers. But this interoperability issues could be solved later when cloud standards are clearly defined and followed by all providers to allow such perfect integration between their technologies and resources. The benefits of using AMGCS are shown in increasing and optimizing the available compute power, managing jobs/resources, and utilizing grid clouds' IaaS resources through integration between system's modules and clouds' APIs. This idea can be beneficial to research centers to solve real-world complex problems that need high computing capabilities, such as Bioinformatics applications, engineering simulations, and mathematical analysis.

Acknowledgment

This Project was funded by Deanship of Scientific Research (DSR), King Abdulaziz University, under grant No. (24/34 Research Group). The Authors, therefore, acknowledge technical and financial support of KAU.

References

[1] Vladislav Falfushinsky, Olena Skarlat, Vadim Tulchinsky, "Cloud computing platform within Grid Infrastructure", Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013 IEEE 7th International Conference on (Volume:02), Sept. 2013.

[2] Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M, "Above the Clouds – A Berkeley View of Cloud", Tech-nical report UCB/EECS-2009-28, EECS Department, Uni-versity of Berkeley, California, 10 February 2009.

[3] Stanoevska-Slabeva, Katarina; Wozniak, Thomas; Ristol, Santi, "Grid and cloud computing: a business perspective on technology and applications", Springer, 2010.

[4] Beaty, Donald, "Cloud computing 101", ASHRAE Journal, Volume 55, Issue 10, p. 88. Oct. 2013.

[5] Jha S, Merzky A, Fox G, "Clouds Provide Grids with Higher-Levels of Abstraction and Explicit Support for Usage Modes". Presentation for Open Grid Forum (OGF) 2008.

[6] José C. Cunha and Omer F. Rana, "Grid Computing: Soft-ware Environments and Tools", ISBN: 978-1-84628-339-0, Springer 2006.

[7] M. Wooldridge, "An Introduction to Multiagent Systems", second ed. John Wiley & Sons, 2009.

[8] K. M. Sim, "Agent-Based Cloud Computing", IEEE Transactions On Services Computing, VOL. 5, NO. 4, December 2012.

[9] A. Lonea, D. Popescu, and O. Prostean, "A survey of management interfaces for eucalyptus cloud," in Applied Computational Intelligence and Informatics (SACI), 7th IEEE International Symposium on, pp. 261–266. May 2012.

[10] X. Wen, G. Gu, Q. Li, Y. Gao, and X. Zhang, "Comparison of open-source cloud management platforms: Openstack and opennebula," in Fuzzy Systems and Knowledge Discovery (FSKD), 9th International Conference on, pp. 2457 –2461. May 2012.

[11] L. Xu and J. Yang, "A management platform for eucalyptusbased iaas," in Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on, Sept. 2011, pp. 193 –197.

[12] C. Baun, M. Kunze, and V. Mauch, "The koala cloud man-ager: Cloud service management the easy way," in Cloud Computing (CLOUD), IEEE International Conference on, pp. 744 –745. July 2011.

[13] C. Baun and M. Kunze, "The KOALA cloud management service: a modern approach for cloud infrastructure management," in Proceedings of the First International Workshop on Cloud Computing Platforms, ser. CloudCP '11. New York, NY, USA: ACM, p. 1:1–1:6. 2011.

[14] "Scalr," [online] Available at: http://github.com/Scalr/. [Accessed: 01 January 2017].

[15] "Apache libcloud," [online] Available at: http://libcloud.apache.org/. [Accessed: 01 January 2017].

[16] "jcloud," [online] Available at: http://www.jclouds.org/. [Accessed: 01 January 2017].

[17] "Apache deltacloud," [online] Available at: http://deltacloud.apache.org/. [Accessed: 19 July 2015].

[18] Yi Wei and M. Brian Blake, "Adaptive Service Workflow Con-figuration and Agent-based Virtual Resource Management in the Cloud", Cloud Engineering (IC2E), IEEE International Conference on, March 2013.

[19] Metsch T., Edmonds. A., et al. Open Cloud Computing Interface Core and Models, Standards Track, no. GFD-R in The Open Grid Forum Document Series, Open Cloud Computing Interface (OCCI) Working Group, Muncie (IN) 2011.

[20] Venticinque S., Tasquier L., Di Martino B., "Agents based Cloud Computing Interface for Resource Provisioning and Management", Sixth International Conference on Complex, Intelligent, and Software Intensive Systems, 2012.

[21] Domenico Talia, "Cloud Computing and Software Agents: Towards Cloud Intelligent Services", WOA, volume 741 of CEUR Workshop Proceedings, page 2-6. CEUR-WS.org, 2011.

[22] ZJ Li, Chen C. and Wang K., "Cloud Computing for Agent-Based Urban Transportation Systems", IEEE Computer Society, 2011.

[23] M.V. Haresh, S. Kalady and V.K. Govindan, "Agent based Dynamic Resource Allocation on Federated Clouds," Proc. IEEE Recent Advances in Intelligent Computational Systems (RAICS'11), pp.111 - 114. 2011.

[24] Del Castillo, Lorenzo and others, "OpenStack Federation in Experimentation Multi-cloud Testbeds." HP Laboratories. 2013.

[25] Kurze, Tobias, et al. "Cloud federation." CLOUD COMPUTING, The Second International Conference on Cloud Computing, GRIDs, and Virtualization. 2011.

[26] Rawat, S. and Rajamani, L., "Experiments with CPU Scheduling Algorithm on a Computational Grid ", IEEE International Advance Computing Conference (IACC 2009), PP. 71-75. 2009.

[27] Chunlin, Li, Zhong Jin Xiu, and Li Layuan. "Resource scheduling with conflicting objectives in grid environments: Model and evaluation." Journal of Network and Computer Applications 32, no. 3: 760-769. 2009.

[28] R. Buyya et al., "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599- 616, June 2009.

[29] K.M. Sim, "Towards Complex Negotiation for Cloud Economy," Proc. Int'l Conf. Advances in Grid and Pervasive Computing (GPC '10), R.S. Chang et al., eds., pp. 395-406, 2010.

[30] K.M. Sim, "Towards Agent-Based Cloud Markets (Position Paper)," Proc. Int'l Conf. E-CASE, and E-Technology, pp. 2571-2573, Jan. 2010.

An Investigation on Information Communication Technology Awareness and Use in Improving Livestock Farming in Southern District, Botswana

Clifford Matsoga Lekopanye
Final Year Student, MSc in Computing,
Teesside University, UK
(Botho University-Campus), Gaborone

Dr. Meenakshi Sundaram K
Faculty of Computing
Botho University
Gaborone, Botswana

***Abstract*—This paper investigated the extent of Information Communication Technology (ICT) usage by livestock keepers and limitations encountered. The study was conducted with the objective of coming up with findings that will contribute towards strengthening ICT usage for the development of livestock keeping. In order to meet this objective the researcher used a mixed method approach where by qualitative and quantitative methods were both used. The results of this study, showed mobile phone technology as the most popular ICT used by 89% of the respondents. Also 73 % of the respondents indicated that they have access to the local radio channel, television accounted for 59%. Other types of ICTs that were pointed out by few respondents are Facebook, Email, the Internet and YouTube. Livestock keepers have identified a number of limitations in using ICTs that need to be addressed that include high cost of communication, poor mobile communication signal, unawareness of television and radio program schedule, and lack of electricity in rural areas.**

In conclusion, this study has identified ICTs such as radio, mobile phones, and television as types of ICT that are used most frequently by livestock keepers, though they are not used at satisfactory level for livestock production. Therefore, the researcher proposed that information systems aimed at delivering information to livestock keepers should be more mobile driven than being computer based. A major approach that could be adopted to address the challenges of radio and television usage is to come up with livestock programs that combine mobile technology with radio and television programs. Participants and listeners to radio or television programs could use mobile technologies to send in their questions either by calls or short messages.

***Keywords*—ICT; Livestock production; ICT utilization; information access; developing countries; ICT awareness**

I. Introduction

Agriculture is a substantial activity in Botswana. The statistics show that 70% of the population is in rural areas and are actively involved in Agricultural activities [1]. Rural resident’s living depends on Agriculture as a source of food, employment, and income. Botswana is a country that generally receives low and unreliable rainfall, regular outbreaks of cattle diseases such as Foot and Mouth Disease (FMD), and recurring droughts.

Popular livestock kept in Botswana are cattle, goats, sheep, donkeys, and poultry. The cattle population in Botswana is 3.6 million. Of these, 15% is under the commercial farming system, and the remaining 85% falls under the non-commercial farming system of small farms [1]. There are two categories of livestock farming systems in Botswana: firstly, the cattle posts, where individual farmers keep small herds of cattle; and secondly the livestock farming that is based on large herds of cattle managed under the communal grazing system [2].

The agriculture sector was contributing about 40% to Botswana’s Gross Domestic Product (GDP) at independence, mainly through the beef exports [3]. The rate has dropped drastically over the years, recording a contribution of 2.6 percent to GDP in 2013. The production was mainly disadvantaged by traditional farming methods, recurring droughts, poor rainfall, and regular outbreaks of pests and diseases [4].

Information Communication Technology (ICT) is an umbrella name that includes any communication device such as computers, telephones, televisions, radio and others. There are also services and software that operates with the aforementioned devices such as email services, livestock management software tools, distance learning software tools and video conferencing applications among others

According to Williams [5], ICT is very important in improving livestock farming. ICT provides livestock farmers with latest livestock farming technology and disseminate formation [5, 6]. Previous researches have shown that ICT is a major contributing factor for improving livestock farming [7]. Local communities or farmers need to take part in decision making during the design and introduction of new ICT systems. Participation of livestock keepers is found to be the most effective way for successful ICT usage in livestock farming [8, 5].

It can be concluded that the use of ICT for record keeping, accessing and disseminating livestock information is an important aspect in increasing profitability. In order for the livestock sector to be economically efficient, livestock keepers need to properly manage all components of the livestock farming [8, 9].

This study will be conducted in order to empirically investigate the ICT awareness and usage by livestock farmers in Sothern district of Botswana with the aim to discover challenges that livestock keepers encounter in ICT usage in

The author would like to thank botho University (Gaborone Campus) for the support and sponsorship of this research work, for publication.

their livestock farming activities. It is anticipated that the research findings and suggested recommendations of this study will improve livestock information access and dissemination in the study area.

II. Research Problem

The problem is, livestock farming productivity have being decreasing in Botswana over the past 50 years. Agriculture sector contributed about 40% to GDP in 1966 mainly through the beef exports [3]. The rate has dropped drastically over the years, to 2.6 percent in 2013, (Honde, 2015). This is a problem because it affects more than half of the population of Botswana. About 70% of the population make their living from farming [1]. Livestock farming is a source of income, source of food and provides employment opportunities for more than half of the population of Botswana. Hence the need to address this problem for improvement of the livelihood of livestock keepers.

There are challenges faced by livestock farmers with regards to livestock information access and dissemination. Lack of ICT infrastructure in rural areas and high cost of ICT services drive livestock keepers to use unproductive traditional ways of farming. They depend on traditional media like television, newspapers and radio as source of information. These traditional mass media cannot provide the services of record keeping nor detect a possible outbreak of diseases, and cannot meet the specific needs of individual livestock keepers. One other challenge faced by most livestock farmers is the language barrier. Most farmers only know their local languages while ICT applications are programed to run in English language. These challenges deprive livestock keepers the benefits that ICT could provide. Effective use of ICT will boost livestock sales and hence improve their livelihood.

According to available literature, no study has been done to determine awareness and usage of ICT for development on livelihoods of livestock keepers in Botswana. It remains unclear whether these ICT challenges apply in Botswana. Hence this study will be conducted in order to empirically investigate the ICT awareness and usage by livestock farmers in southern district of Botswana.

III. Purpose of the Study

Most of the livestock farmers are elders and their farming methods depend on traditional knowledge or local knowledge. Information knowledge, management, and communication plays an important role in development of livestock farming. With absence of ICT intervention or ICT usage farmers may remain using their old unproductive traditional knowledge of livestock farming. This study will be carried out in order to suggest ways that can improve livestock information access and dissemination.

According to available literature, no study has been done to determine awareness and usage of ICT for development on livelihoods of livestock keepers in Botswana. A comprehensive understanding of the relationship between ICTs and livestock keeping is still lacking in Botswana. Therefore this research is expected to bring about knowledge and awareness of the challenges encountered on usage of ICTs by livestock keepers. Challenges that apply in other countries may not apply in Botswana. It is therefore, anticipated that this study will add new knowledge in ICT and livestock farming in Botswana. This study is conducted with the purpose of coming up with findings that will contribute towards strengthening ICT usage in livestock keeping. This study will be carried out to determine the level of ICT awareness and usage by livestock keepers. The study is meant to highlight the gap that exists on the ICT development in livestock sector in Botswana. The question is: Do livestock keepers in Botswana use ICT? If the answer is ‘yes’, then, to what extend are they using it? If the answer is ‘no’, then, what are the limiting factors? Therefore the study is designed to seek for answers to these questions.

IV. Research Objectives

To find the level of ICT awareness and ICT usage by livestock Keepers in southern district of Botswana with the aim to discover challenges that livestock keepers are facing in using ICTs in livestock farming, in order to suggest ways of improving their ICT usage in their livestock farming activities.

A. *Specific Objectives*

1) To identify types of ICTs that livestock keepers use in the study area.

2) To examine the usefulness of ICTs used on accessing and disseminating livestock information.

3) To determine the economic status of livestock farmers in relation to the use of ICT.

4) To identify challenges encountered by livestock keepers in using ICTs in order to suggest ways of improving information access and dissemination.

V. Research Questions

A. *Main question*

To what extent are livestock keepers aware and use ICT in the study area?

B. *Specific questions*

1) From the existing types of ICTs, which ones are known and used by livestock keepers in the study area?

2) Are livestock keepers aware of theimportanceof ICT usage as information sharing tool in their livestock production business?

3) How does the economic status of livestock keepers relate with their ICT usage?

4) What limitations do livestock keepers encounter in using ICT?

VI. Literature Review

This study investigates ICT awareness and usage for livestock farming development in the southern rural community of Botswana. The basis for conducting the study is to investigate on the limitations or challenges met by livestock keepers in accessing and using ICTs in rural communities. These limitations have negative effect in their productivity and the volume of their livestock production. The fundamental argument of the study is that employment of ICTs on livestock keeping has positive results of improving the quality and quantity of livestock production. In addition lack of ICT awareness can result in underutilization of available ICTs that could have enhanced livestock productivity.

A. *Types of ICTs that are known and used by livestock keepers in the study area*

The role of ICT in livestock farming is defined or determined by farmers, i.e. their ability to adopt and effectively use ICT in livestock farming activities. Use of ICT enables easy access and exchange of information which in turn increase efficiency, competitiveness and productivity in various livestock farming activities [7].

According to Angello [6] large scale commercial livestock keepers are expected to use digital imaging, cameras, computing devices, the internet, Wi-Fi, and Wireless Access Protocol (WAP) based Internet access among others, while small scale farmers utilize basic types of ICT such as computers, Internet and mobile phones.

Of all technologies, mobile phones are certainly the technology of choice for many livestock keepers in both large and small scale livestock production [7],[10], [11]. Mobile phones are cheap, easy to manage power efficient and have internet access that enable farmers to compare prices more efficiently and helps to improve the links between farmers and the market, creating more chances for small scale farmers to make more sales and hence improve their livelihoods.

The study that was conducted in Tanzania by [10] indicated that mobile phones were used most frequently by livestock keepers in getting information concerning their business. Other types of ICT such as radio and television are also used by a significant number of Tanzania livestock keepers. The Internet is accessed and used by quiet a small number of livestock keepers. Other studies that were conducted by [11], [7] showed similar results of high use of mobile phones, followed by television and radio with a minimum use of Internet, computers and email by livestock keepers.

B. *Awareness of the importance of ICT usage by livestock keepers*

The study that was done by [10], discovered that livestock keepers were aware of the importance of ICT usage.Furthermore, they indicated the need for information dissemination in order for their livestock keeping activity to grow and benefit them become more profitable. The study showed that 80% of the livestock keepers complained about lack of training personnel and also there is no time to conduct trainings. Livestock keepers suggested the use of television and radio to inform and educate livestock keepers since they have a wide coverage. According to Temba [10] livestock keepers requested that the broadcasting media should disseminate information on livestock diseases and control. From this, one can conclude that livestock keepers see the importance of ICTs such as radio and television. Therefore it is left for the broadcasting media to develop educative and informative programs on livestock keeping.

From the study that was conducted by [6], all the respondents that were part of the study did not have websites. Participants were not aware of websites, and a few reported to be using emails and livestock management applications. It is not surprising to get this kind of results more so that most livestock keepers are not computer literate. Livestock keeping applications and websites are more advanced information technology as compared to simple and cheap mobile phone technology that is mainly used for making calls and short messages.

C. *The economic status of livestock keepers in relation to their ICT usage*

One thing that can be used to measure how ICT usage relates with the socio-economic status of livestock keepers is the usage rate and diffusion rate of mobile phones. In Botswana the use or dispersal of mobile phones grew exponentially from 3301 in 1998 to 3,460,331 2016 [13]. One of the reasons for this is the low cost of mobile phones. The cost of a low specification mobile phone is about BWP 200.00 (US$ 18.94). This shows that livestock keepers in Botswana are ready to adopt ICT when affordable. The text message service is the most popular service which clearly shows that most people can use related ICT services such as emails if they have access and can afford it.

For farmers to rip the benefits of ICT, first there should be ICT resources in place such as the computer hardware, computer software and internet access. However, the study that was conducted by [12] showed that most of livestock keepers in Botswana cannot afford to have ICT resources and internet access due to high costs. The cost of a minimum requirements specification computer is about BWP 3, 500.00 (US$ 331.53) and internet connection charge is about BWP 500.00 (US$ 47.36)

D. *Challenges that livestock keepers encounter in using ICT*

According to Arshad [14] livestock keeping can be improved through educating or training livestock keepers in their particular field. It is not easy and is expensive to conduct such trainings due to poor roads and dispersed livestock keepers. ICTs such as video conferencing, television and radio broadcasting could be used to educate livestock keepers at their particular field. However, these ICTs come with their challenges such as high cost of ICTs, lack of skills to use or operate them [14] and others that are discussed below.

1) *High cost of ICT and Electricity connection*

According to Mooketsi[15] few livestock keepers in Botswana use Internet in order to access online services due to lack of ICT infrastructure and electricity connection in rural areas. However, the author has observed that Batswana livestock keepers meet their goals when they work in collaborations. Livestock keepers in a certain area will come together and form syndicate in order to share the costs of drilling boreholes [15]. From this the author suggested that livestock keepers could possibly use the same idea to form groups in order to share costs of buying computers, connecting electricity to their homes or farms, and paying for the Internet cost. Therefore, livestock keepers are encouraged to take such initiatives and enhance their ICT usage in livestock keeping.

2) *Poor ICT infrastructure in rural areas*

The government of Botswana has been doing some developments over the past ten years in providing relevant ICT infrastructure. A lot of financial resources have been spent in

ICT infrastructure. For the period of 2003 to 2010 the government invested 3.7% of national budget [16]. However, most of these ICT developments were implemented in government institutions such as secondary schools. Other ICT infrastructure challenges such as poor or limited radio and television signal coverage, Internet and power connection are not addressed, especially in rural communities.

3) ICT Initiatives that are web-based

Existing literature discuss some of the initiatives that employ the use of web-based ICTs in livestock keeping. This includes the VERCON in Egypt that use the Internet to establish and strengthen linkages among the livestock keepers, extension officers and agricultural researchers [17]. The VERCON connects geographically separate people with an enhanced two-way communication mode. Another initiatives that [18] has written a report about is the Linking Local Learners (LLL) in Kenya. This is a virtual network that connects livestock keepers' syndicates' together. In Uganda they have set up the District Agricultural Training and Information Centres (DATICS). In South Africa, a centralized Integrated Registration and Generic Information System (INTERGIS).

This projects or initiatives that are to fill the ICT gap in the livestock sector,however most of them are web-based. This makes it difficult for illiterate farmers to access and disseminate livestock information. Also, web-based solution brings about another challenge of poor Internet infrastructure and where it is available; it turns to be very expensive. Since, these initiative solutions are very useful, what can be done is to come up with easy ways for livestock keepers to navigate through, even those with low level education. The use of mobile phones is encouraged as they offer easy accessibility and are cheap to obtain. Although the use of short messages is cheap and requires basic level of literacy, it carries only a limited amount of information.

VII. METHODOLOGY

A. The research area

This research study was conducted in the Ngwaketse Region, located in southern District part of Botswana. The Region lies between latitudes -25° South of the Equator and between longitudes 25° East of Greenwich. The Ngwaketse region comprises of five (5) sub-regions, namely Barolong, Ngwaketse North, Ngwaketse South, Ngwaketse Central, and Ngwaketse East regions. This districts was selected purposively because of its bigger size, it has a high number of livestock holdings of 13, 202. According to CSO annual agricultural report, Sothern District has a total of 263, 974 cattle. The region is the second largest area to host beef farmers. There are several beef ranges that are run by the government. Also, there are privately owned large ranges that support livestock keepers in the area[1].

B. Research design, sample size and sampling techniques

In order to get answers of the research questions, the researcher adopted a mixed method approach. This means that quantitative and qualitative methods were both used. The researcher used observations and interviews for collecting qualitative data. The qualitative data was use to confirm the results from the collected quantitative data. Questionnaires were used to collect quantitative data so that the findings could be quantified in terms of percentages and frequencies. According to Creswell [19] it is advisable to use a mixed method approach because one of the used method will neutralize the shortfalls of the other used method.

In this study, both purposive method and snowball sampling techniques of non-probability were used in selecting the study sample. A combination of two non-probabilities. The purposive or judgmental method was used by the researcher to identify respondents that were most likely to answer the questionnaire and interview questions in a way that will lead to attaining the objectives of the study. Ministry of Agriculture have placed extension officers in rural areas who played a big role in helping the researcher to select more informative respondents. The researcher also, purposively selected areas with high density of livestock keepers in order to increase the response rate.

Snowball, was used to select the participants. This method is used when it is not easy to identify the respondents or the participants of the study [20]. With Snowball sampling the researcher has to attend one or two respondents in order to identify other respondents. In this study, one livestock keeper was interviewed and thereafter identified other livestock keepers who also helped to identify other participants until the required number of participants is met.

The target respondents for this study were livestock keepers in the southern district part of Botswana. Statistical data of livestock keepers in the study area was extracted from the Central Statistics Office Annual Report of 2014 and shown in Table I. The number of livestock holdings form each sub-region was determined by the number of livestock farms. In order to get the sample size for each sub-region, the researcher used a proportional to size sampling method. The sample size was calculated using the formula:

$$Sample\_Size = \frac{a}{b} * c \qquad (1)$$

Where 'a' is total number of livestock holdings in a sub-region, 'b'is the total number of livestock farms in the hole region or district, and c is the total sample size in the region, i.e. 'c'= 60.

TABLE I. CATTLE FARMS PER SUB-REGION IN SOUTHERN DISTRICT

Sub Region	No. Of livestock Holdings	Sample Size
Barolong	1,775	8
Ngwaketse North	1,389	6
Ngwaketse South	1,379	7
Ngwaketse Central	4,487	21
Ngwaketse West	2,172	10
Total	**11, 202**	**60**

C. Data collection techniques and analysis

In this study, primary and secondary data was collected with the survey method. Primary data was collected using observations, interviews and questionnaires. Secondary data was collected from several different sources such government surveys, media, various publications such as books, journals and Internet material. Collected data was systematically recorded and analysed.

VIII. Results and Analysis

A. Socio-demographic data

Table II shows the socio-demographic data of 52 respondents from the 60 that was targeted. There are 60% of male livestock keepers while 40% are females. Most livestock keepers, 52% are aged between 30 and 49, while respondents aged above 49 accounted for 31%, and livestock keepers in the age group of 18 to 29 accounted for 17%.

Fifty three percent (53%) of livestock keepers have been in this sector for 15 years, and few respondents (6%) have been in livestock keeping sector within the past 5 years. Respondents with primary education have the highest percentage (42%), followed by those with secondary qualification (27%), 12% attended tertiary education, 12% did adult education and 8% have no School qualification at all.

TABLE II. Socio Demographic Data

Variable	Measure	frequency	percentage
Gender	Female	21	40%
	Male	31	60%
Age	18 to 29	9	17%
	30 to 49	27	52%
	above 49	16	31%
Farming Experience	1 to 5	3	6%
	6 to 10	8	15%
	11 to 15	12	23%
	above 15	29	56%
Education	Adult education	6	12%
	Primary Education	22	42%
	Secondary Education	14	27%
	Tertiary Education	6	12%
	Never in School	4	8%

B. Level of ICT usage by livestock keepers

In one of the research questions the respondents were to declare whether they use ICT for their livestock production activities. Quiet a large number of the respondents (87 %) indicated that they usedat least one of the ICTs for communicating livestock information while 7 respondents (13 %) indicated that they do not use any of the ICTs. The results of this question are shown in Fig 1.

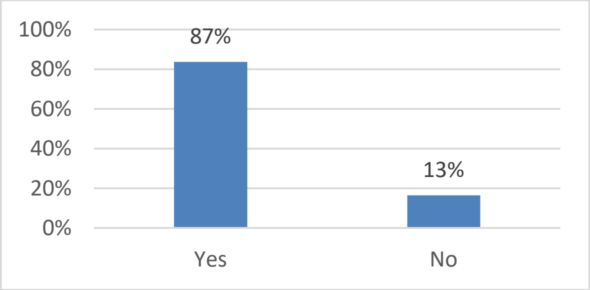


Fig. 1. Level of ICT usage by livestock keepers

The compilation of the survey showed that ICTs are used by many livestock farmers in southern district of Botswana. This is a clear sign that proper use of ICTs can improve livelihoods oflivestock keepers because information shearing between livestock keepers themselves, extension officers and veterinarians may lead to improved livestock keeping practice and quick solving of encountered problems.

Furthermore, the respondents were to indicate the types of ICTs that they use. It appears that most livestock keepers use the radio for accessing information and learning, some use local television, mobile phones, and few livestock keepers have Internet access.

The results displayed in Fig 2. show that mobile phone technology is the most popular ICT used by 94% the respondents. Livestock keepers also indicated that they have access to the local radio charnel, and more than half of them can afford to buy at least a small FM Radio. Respondents who listen to the radio accounted for 71%, television accounted for 40%. Another type of ICT that was pointed out was the use of Internet by 21% of the respondents. The results show that the respondents use internet for Facebook (17%), YouTube (4%), and email (8%).

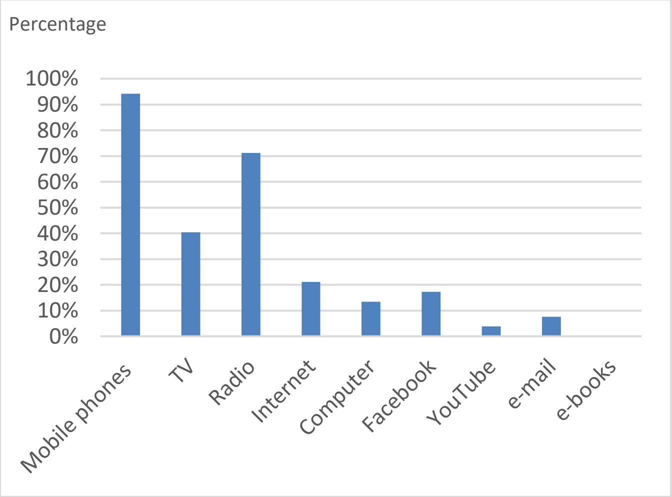


Fig. 2. Typesof ICTs that are used in the study area

C. Importance of ICTs in livestock information dissemination

Sometimes the ICTs are not effectively utilized because users don’t see the importance of using ICTs in their business. The researcher wanted to understand whether the respondents are aware of the importance of ICT tools to them in accessing livestock information. The findings of this study have shown that a high percentage (87%) of respondents agreed that ICT usage is important to them for accessing livestock information. A small percentage of respondents (13%) do not see the importance of ICTs.

The reasons given by the respondents were summarized by the researcher as presented in Table IV. Most of the respondents (35%) found ICT important in making communication easier and fast. According to my observation and discussion with the respondent all of these respondents use mobile phones and they gave different reasons on how mobile phones make communication easier. Some use mobile phones to communicate with other livestock keepers; others make phone calls to veterinary doctors, and this result in getting help in time. Only 31% of the respondents indicated that ICT is important since they use it to get technical advice. The respondents get the advice through the use of Radio, Television, computers and Internet.

TABLE III. IMPORTANCE OF ICTS IN LIVESTOCK INFORMATION DISSERTATION

Importance	Frequency	Percentage
Yes I get technical advise	16	31%
Yes it makes communication easier	18	35%
Yes But I have no time to use ICT in livestock keeping	11	21%
I **don't** see the importance in using ICT in livestock keeping	7	13%

D. Limitations of using ICTs in the study area

The outcomes of the question that wanted the farmers to state the factors that hinders them from using ICTs for information dissemination, revealed a number of factors that need to be addressed. Table V presents the summery of the response from the participants.

TABLE IV. LIMITATIONS OF USING ICTS IN LIVESTOCK KEEPING

Limitation	Frequency	Percentage
High communication costs	44	85%
Poor network	39	75%
Long distance to Internet services	17	33%
Lack of computer skills	36	69%
Unawareness of the programme schedule on Radio/Television	7	13%
Not aware of programs	11	21%
No electricity	28	54%
Lack of confidence in operating ICTs	22	42%
Insufficient regional specific language	14	27%
Negative attitude towards ICTs	4	8%
Lack of training	41	79%

1) High cost of communication

Almost all respondents mentioned that the cost of communication with mobile phones is high. The mobile phone call rates are said to be high especially during business hours. The respondents also complained about the high cost of internet subscription, stating that this high communication charges are affecting use of mobile phones. The respondents mentioned that the Government does not give allowances for mobile communication to the extension officers. They use their personal mobile phones at their own expenses and this hampers the use of mobile phones in information dissemination since extension officers who have knowledge in livestock farming cannot communicate with all livestock keepers.

2) Poor mobile communication Network signal

Another limiting factor that was mentioned by 75% of the respondents was poor mobile communication Network signal. In addition, the researcher observed that in some areas there is no Network signal at all. Farmers stated that sometimes they have to travel to nearby areas where there is a communication Network In order for them to use their mobile phones. This increase the communication cost and time consuming.

3) Unawareness of radio and television programmes and their schedules

Some respondents, 13% are not aware of the time programs were broadcasted. This is a limitation because respondents cannot listen or follow this informative program. Even the worst case, some livestock keepers 11% do not know that there exist educative livestock programs on radio and television. These are the livestock keepers who declared that they do not have radio nor television set at their homes due to different reasons such as: lack of money to buy; poor or no radio and television signal in their area; lack of electricity and others.

4) Lack of electricity

This study reports that 54% of the respondents have no electric power connection in their homes. The researcher observed that power connection is not yet done in rural areas that are distanced from the main tarred roads. Power connection is done in livestock farms and homes alongside the main roads. The respondents said that this situation is depriving them the chance to watch television and listening to the radio in order to benefit from these livestock programs. Some livestock keepers who have electricity complained about the national crisis of power rationing.

IX. DISCUSSION OF FINDINGS AND RECOMMENDATIONS ON IMPROVEMENT OF ICT USAGE

This study is focused on finding out the level of ICT awareness and usage by livestock farmers in Southern District part of Botswana. The researcher want to understand the types of ICTs that are used, and if livestock keepers are aware of the importance of ICT usage in their livestock farming activities. Data was collected from 122 livestock farmers in the study area. A quick observation as well as secondary evidence from Central Statistics Office Annual report (2014) shows that goats, cattle, sheep and poultry are the major livestock kept by livestock keepers in the study area.

During data collection, the researcher had to first explain to most of the participants why they need to participate in the study and the question had to be translated and explained to them in their local language, Setswana. This was an exceptional to a few large farms at Borolong area. The few large farms had more staff; some had offices unlike the small scale. Some of the large farms used diesel generators for electricity.

The percentage of male livestock keepers (66%) who participated is higher than that of female livestock keepers (34%). Most farmers are aged between 30 and 49 years old. The study indicates the participation of young persons. Remarkably more than 33% reported that they have secondary and tertiary education.

Almost all the participants can afford and are using mobile phones. Also about six out of ten have access and are using radio and television. From the literature review of this study, a number of studies that have been conducted in some African countries had this pattern of response.

It turns out that mobile phone technology is currently the most popular ICT of the present society. This is because of its portability, low cost, simple to use together with its capabilities of multitasking. Livestock application designers and developers are therefore, advised to come up with systems that operates on commonly available technology devices, i.e. mobile devices for now. Also the systems need to be presented in local languages in order to include the largest possible number of end users.

Apart from mobile phones, it is also observed that radio is listened by a large percentage of livestock keepers. This is because radios have unique qualities, and most mobile phones have radios embedded making the radio to be one of the popularly used ICT. Radios can operate on batteries, are affordable, and most mobile phone have FM radio embedded. However, to my observation and discussion with the respondents, there is a lack of awareness that their mobile phones have radios embedded. One of the short falls of the radio is that it is a one way communication system unlike mobile phones. For this radio presenters need to be clear and explain in details to ensure that listeners have no questions in the information that they receive.

In regardless of the good qualities of the radio, the results of this study showed that there are livestock farmers who do not know about the existence of radio nor livestock programs. The radio and television programs' details need to be advertised through other mediums such as newspapers, and during gatherings such as 'Kgotla meetings'. This can increase the number of livestock keepers who follow or listen to them and hence benefiting from them.

There are livestock keepers who are not able to follow the television and radio livestock programs because of unfavorable time schedule of the programs. Livestock keepers indicated that programs are scheduled on day time when livestock keepers are out doing livestock rearing activities. As a result, it is advisable to review and improve these programme schedules in order to attract more listeners. The researcher suggests that the programmes should be broadcasted at night around 08:00 pm when most livestock keepers are available to watch and listen to this program. Also, livestock programs need to be made more interactive by inviting the listeners to make phone calls to the radio and television to seek for clarification and ask question. This can be possible since a lot of livestock keepers since almost all of them have mobile phones.

X. Conclusion

This study has discovered that use of ICTs by livestock keepers for the purpose of improving their livestock production activities in the southern district of Botswana is not satisfying. Radio, mobile phones, and television are the identified types of ICT that are used most frequently by livestock keepers, though they are not used at satisfactory level for livestock production. ICT usage and access is limited by lack of computer skills, high cost of computers and Internet access, unawareness of livestock programs on radio and television, lack of ICT infrastructure in rural areas, and others. If these challenges can be resolved, there can be a significant improvement in livestock production especially in remote rural areas where livestock keeping is highly practiced.

References

[1] CSO, "Annual Agricultural Report," Central Statistics Office, Gaborone, 2014.

[2] D. B. Gosalamang, J. J. Hlongwane and M. Masuku, "Supply Response of Beef Farmers in Botswana: A Nerlovian partial adjustments Model approach," African Journal of Agricultural Research, vol. 7, no. 31, pp. 4383-4389, 2012.

[3] B. Malema, "Botswana's formal economic structure as a possible source of poverty: Are there any policies out of this economic impasse?", PULA: Botswana Journal of African Studies, vol. 26, no. 46, 2012.

[4] G. Honde, "Botswana," AfDB, OECD, UNDP, 2015.

[5] E. E. Williams and I. S. Agbo, "Evaluation of the Use of ICT in Agricultural Technology Delivery to farmers in Ebonyi State, Nigeria," Journal of Information Engineering and Applications, vol. 3, no. 10, 2013.

[6] C. Angello, "Exploring the Use of ICT in Learning and Disseminating Livestock Husbandry knowledge to urban and per-urban communities in Tanzania," International Journal of Education and Development Using Information and communication Technology(IJEDICT), vol. 11, no. 2, pp. 5-22, 2015.

[7] N. Williams and O. Soremi, "ICT Use in Livestock Innovation Chain in Ibadan City in Nigeria," Advances in Life Science and Technology, vol. 32, pp. 30-43, 2015.

[8] B. Rudolph , "Cattle Traceability-A Threat to Sustainable Suppy of Beef to EU: A Botswana Meat Commission," European Centre for Research Training and Development UK, vol. 1, pp. 1-9, 2013.

[9] M. Meyn, "The end of Botswana beef exports to European Union," London SE 17 JD: Overseas Development Institute, 111 Westminister Bridge Road, 2007.

[10] B. A. Temba, F. K. Kajuna, G. S. Pango and R. Benard, "Accessibility and use of information and communication tools among farmers for improving chicken production in Morogo municipality, Tanzania," Livestock Research for rural development, vol. 28, no. 11, 2016.

[11] W. P. Mtenga and A. C. Msungu, "Using Information and Communication Technologies for enhancing the accessibility of agricultural information for improved agricultural production in Tanzania.," The Electronic Journal on Information Systems in Developing Countries, vol. 56, no. 1, pp. 1-14, 2013.

[12] T. Mogotlhwane, F. Khosrowshahi and J. Underwood, "ICT Challenges in Developing Countries: Botswana's Perspective," International Journal of Computer and Information Technology, vol. 2, no. 6, pp. 1054-1058, November 2013.

[13] BOCRA, "BOCRA," 15 November 2016. [Online]. Available: http://www.bocra.org.bw/telecoms-statistics. [Accessed 17 November 2015].

[14] S. Arshad, A. Saghir and M. Ashfaq , "Gender and Decision Making Process in Livestock Management," Sarhad J. Agric, vol. 26, no. 4, pp. 132-135, 2010.

[15] B. E. Mooketsi, "Optimization of Livestock Identification and Trace-back System LITS Database to Meet Local Needs: Case Study of Botswana," The Journal of Community Informatics, vol. 9, no. 4, 2013.

[16] T. Mogotlhwane , F. Khosrowshahi and J. Underwood, "Information technology productivity paradox," in 6th International Postgraduate Research Conference, Netherlands, 2006.

[17] Case study: Institution-Based Information System, Egypt, The Experience of VERCON in Egypt, 2016.

[18] Marketing: Linking local learners, 2016.

[19] J. W. Creswell, Research Design: Qualitative, Quantitative and Mixed Methods, Second ed., SAGE Publications, 2003.

[20] M. Saunders, P. Lewis and D. Thornhill, Research Methods for Business Students, London: Pearson Education Limited, 2007.

Tutoring Functions in a Blended Learning System: Case of Specialized French Teaching

Nadia Chafiq
Member of the multidisciplinary Laboratory in Sciences and Information, Communication, and Education Technology (LAPSTICE)
Observatory of Research in Didactics and University Pedagogy (ORDIPU)
Faculty of Sciences Ben M'Sik, University Hassan II of Casablanca, Morocco

Mohammed Talbi
Laboratory of Analytical Chemistry and Physical Chemistry of Materials
Observatory of Research in Didactics and University Pedagogy (ORDIPU)
Faculty of Sciences Ben M'Sik, University Hassan II of Casablanca, Morocco

***Abstract*—There is an emergence of blended learning today which combines diversified teaching methods, alternating distance learning and classroom learning. As a matter of fact, most Moroccan universities are presently aware of the importance of this approach, which appears to be most suited for Moroccan university context. This article is meant to identify the different roles of the tutor within the blended learning system. This is more precisely to present an experience of implementing a hybrid learning system by using the "FOUL" platform. The introduction of this platform is accompanied by a need for developing new skills, be it for a teacher or a student. This experience is motivated, on the one hand, by the supply of additional online resources as a complement to a face-to-face classroom method , on the other hand by the personalization of learning and riding out classroom-based learning constraints (e.g. in terms of time, place, staff ...) as it is the case in universities. This article intends to address the above problems by analyzing student responses to questionnaires and processing the content of synchronous communication between learners/learners and learners/online tutors in order to identify and analyze tutoring functions. It can be concluded that the success of a hybrid learning system is conditioned by the presence of some basic functions such as: pedagogical, organizational and socio-motivational functions. These functions remain dominant in a hybrid learning system.**

***Keywords*—Learning scenario; Tutoring functions; platform; linguistic proficiency; Interaction**

I. Introduction

The sociolinguistic question is one of the issues in Morocco. At the level of secondary education, the scientific disciplines are being taught in Arabic. The Arabization is not carried out at the level of university and the students are dealing with scientific courses in French and so facing serious linguistic difficulties. Since its implementation in universities with an open access (2003), the teaching of language and communication has been the object of many pedagogical reforms tempting to make it more coherent and better adapted to the real needs of students. The last reform (2014) renames the module ‘’language and Communication (LC)’’ which is, currently, entitled ‘’Language and Terminology’’. The latter is carried out in the first and second semester with a timetable of 45 hours each shared between the teachers in classroom and online via the Moodle platform. Therefore, a new system referred to as‘’ Hybrid System’’ is being established.

The teaching of the French language in Moroccan universities adopt the spirit of the Common European Frame of Reference for languages (CEFRL) whereof its existence is motivated by the necessity: of a common basis for the making of language programs, the design of exams and textbooks ; of a descriptive frame to delineate the learning objectives of a language so as to use it in order to communicate ; to set down the knowledge and the skills to acquire in order to possess an efficient linguistic behavior (CEFRL, 2001, p.9).

To start with, the teaching of the module Language and Terminology (L/T) in The Moroccan University has been particularly designed for a large and not highly motivated arabized turnout, in a hurry and with few resources. The computing tool has been cogently imposed as a supplement meant to improve this situation. To begin with, once the context and issues of this research are laid out, the Faculty of Sciences Ben M'Sik student’s responses to questionnaires will be analyzed. Then the main constraints related to the hybrid system and the tutoring functions of the teacher will be defined. Finally, the results will be analyzed and discussed in order to identify the key factors to make the hybrid system operational within the Moroccan university.

II. Context

The Moroccan university receive a "massive number of students", but beyond the quantitative aspect of the phenomenon, it is important to take into account that at the same time as this "mass" grows, it also diversifies and this is precisely what makes it necessary, and even indispensable, to implement training systems that can meet the wide-ranging needs of current learners. Therefore, training systems must "no longer be what they had been" and then should adapt to the needs of students. The approaches that offer the most autonomy to the student, and which are highly praised: hybrid approach, workshops and language laboratory.

In Morocco, as in other countries, teachers seem reluctant to exploit such systems despite the fact that much effort has been made to further promote the development of e-Learning and to make the hybrid system operational within the Moroccan university so that it would be on the same footing as the international universities. The choice of this system will grant Moroccan students a latent period by continuing to offer hybrid trainings, alternating face-to-face and distance classes, before switching to the 'all on line'.

By working in a network or alone on the computer, some students who are not used to learning autonomously would be tempted to abandon the program at the slightest difficulty. It is therefore appropriate within the framework of the hybrid system to involve trainers capable of designing well-structured programs enabling learners to exploit their time efficiently. Online tutors must also be competent to better support students' commitment.

III. Issue

Among the most important constraints facing the implementation of the hybrid system at present are those of an organizational and pedagogical nature. This paper argues that if the hybrid system is associated with supreme supervision and mentoring, it will be able to overcome these constraints and help students develop their language and communicative competence.

However, the tutor, as competent as they may be, cannot be effective, without setting up normative systems regulating the e-learning project, and including formal elements such as laws, structures (e.g. a language centre), The number of tutors should, for example, be proportional to the number of learners, the calculation of the teaching hours which behoove the tutor at the e-Learning level, the reward systems, etc. A prerequisite is therefore to institute an ethical organizational climate for the e-learning project that influences the tutors' decision-making. Tutoring is often the last thing designers of learning systems are interested in, only concerned with, technological solutions.

Thus, questioning tutoring raises several questions about the identity of tutors, their roles, the planning of their actions, and so on. Moreover, in a context of hybrid training, the tutor's function is diversified and renewed, which necessitates a reflection on a descriptive model of the tutors' functions. For example, it is important to question the interventions at distant but also face-to-face classes and how they can be complementary. In addition to their traditional pedagogical objective linked to the transmission of knowledge, tutors must be able to tackle the technological, organizational, socio-affective, relational and metacognitive problems that punctuate the activities of the learners: What are the fields of intervention of the tutor in the context of a hybrid system? What are the tutorial functions of the language tutor (Analysis of participatory feedbacks-Moodle)? What new supportive skills are needed for these tutors within the hybrid system?

IV. Methodological Choice

A. *Techno-pedagogical environment*

A sample of 120 physics students (Level A2), were the subject of this research. (see table below).

TABLE I. Experiment / Hybrid System

Platform	Number of tutors	Number of students	Public	Face-to-face teaching	Remote teaching
FOUL (2014/2015)	4 tutors	120	A2 level S1	25H (Cap university handbook/ environment)	15H 3 pedagogical online scenarios

"FOUL" (French for University Objectives) is the suggested platform for experimentation at the Faculty of Sciences Ben M'Sik. This platform was created with Moodle (http://tice-lt.info/foul/). Then the content of the platform was based on the results of an analysis of student's needs. FOUL consists of thirteen units corresponding to thirteen different themes related to the syllabus of the first year of the bachelor degree, and they are: The elements of nature, shapes and colors, Scientific press, the digestive system, the influence of light on plants, the immune system, GMOs, nuclear waste management, IT security, renewable energy, nanotechnology, the forest ecosystem and application form.

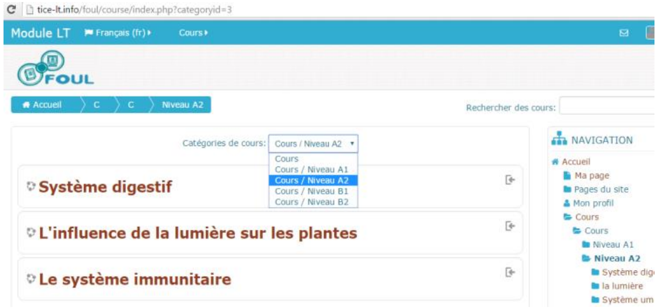


Fig. 1. Example of a figure caption - FOUL Platform

Each of the units is then divided into six sections: identification of the scenario, awareness, exposure, appropriation, assimilation / production and evaluation (of a project). The first three are a set of micro-tasks, in accordance with action-oriented approach - or approach tasks - recommended in 2001 by the European Framework of Reference for Languages (CEFR now), help to achieve a macro final-task, which is the "educational project" in the case of FOUL.

Video documents, texts, images and sound are used to trigger the achievement of different activities. Finally, this training involves the presence of a tutor, which must take place in the various Moodle tools used by students, namely e-mail, chat, forum and at the correction of the activities. This tool has been integrated in our educational system to develop the hybrid scenario and promote the acquisition of new knowledge and skills by students outside the classroom. The online courses will be complemented by the French handbooks "Cap university" that address A2 level students, as classified by the Common European Framework for Languages (CEFR) and aims to get them to B1 level.

First, the investigatory work done was based on online survey on 120 students to collect information on the use of tools of the platform, their perceptions of mentoring and online learning difficulties. Furthermore, to better understand the diversity of situations actually experienced by students during the hybrid system and to avoid the limitations of the "quantitative" survey. Interviews have been conducted and the content of synchronous communications between students / learners and learners / online tutors have been analyzed. This helped in reflecting upon tutorship functions

B. *Presentation of the model used in research*

Implementing hybrid systems is a complex process. On one hand, it takes into consideration several aspects:

The organization of the training, roles of involved parties, the lesson plan and resources. On the other hand, it is part of a dynamic process of adapting to changes in society, technology and training paths. Making the choice of a hybrid system that implies changes in terms of organization, teaching and learning, hence the need to think about modeling. According to Gilles Willett [1]"there is a tendency to view the models as being first in the schematic representation to describe and illustrate reductively, simply and functionally the essential features of an object, of a system or a process."

The models are used to create some order between the elements of a complex whole and represent the links, the connections and the relationships between these elements. No model can be applied at all levels of analysis and all research objectives. The model used in this research sets, reports and describes the tutorship functions as part of a hybrid system (see diagram below).

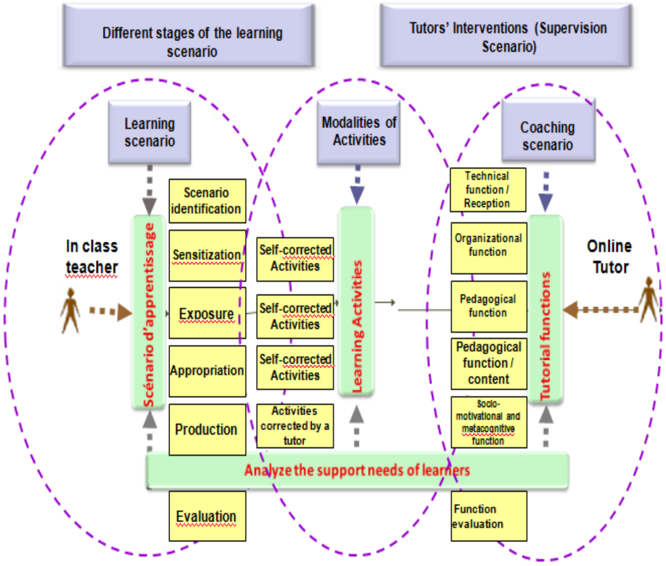

Different stages of the learning scenario
Tutors' Interventions (Supervision Scenario)
Learning scenario
Modalities of Activities
Coaching scenario
Scenario identification
Sensitization
Exposure
Appropriation
Production
Evaluation
Self-corrected Activities
Self-corrected Activities
Self-corrected Activities
Activities corrected by a tutor
Technical function / Reception
Organizational function
Pedagogical function
Pedagogical function / content
Socio-motivational and metacognitive function
Function evaluation
In class teacher
Online Tutor
Scénario d'apprentissage
Learning Activities
Tutorial functions
Analyze the support needs of learners

Fig. 2. Descriptive Model of tutorship functions as part of a hybrid system

The model describes the distribution of roles of remote tutors to face-to-face tutors, the functions of supervision and modalities of learning activities. Indeed, the hybrid system requires from the teacher a rigorous training organization and a clear definition of remote tutor intervention methods and face-to-face teaching. It is a complex system that requires anticipation of actions related to both the field of supervision and that of learning. Indeed, the anticipation of educational activities refers to the notion of the pedagogical scenario and it comes in two forms, the learning scenario and the supervision scenario (Quintin, JJ &Depover, C. &Degache, C. (2005))[2].

At the platform level (FOUL), a learning scenario will be described as a series of steps which are: identification of the scenario, awareness, exposure, ownership, production and evaluation (of a project). As for the supervision scenario, according to the model above, it defines the tutorship functions of the language tutor in the learning process.

V. Analysis of Results

A. *Analysis of the tools of the tutor*

The platform FOUL provides tools to help the tutor in his coaching task: consulting tools to give access to knowledge (links, documents, books) but also tools of exchange (Forum, chat, assignment, wiki, and logbook) to support the construction of knowledge. The table below demonstrates the analysis of two of the tutor tools: Forum and Chat.

TABLE II. Tutor Tools / Forum and Chat

Platform	What is a forum?	What is Chatting?
To Pass on information ?	YES- Careful of the loss of information	YES- in the case of a Initial chatting
Communicate and interact ?	yes	Yes
Follow the activity of learners ?	yes	Difficult
Co-creation of content ?	yes	Possible with wiki
Evaluate the learners ?	yes	Quite difficult especially in languages

In the student survey, students were asked to give their opinion on the use of tools of the FOUL platform (Chat and Forum). For chatting, 50% of students consider this tool useful for team meetings. Indeed, in general, students use the chat tool to ask questions on the course to online peers or to contact the teacher.

However, there have not been many exchanges. Initially, there was willingness on the part of the tutors, it was originally planned that this platform will be used as a tool for exchange but it was noticed that there were some attempts; only a few words, perhaps in two or three sessions, were observed. The lack of communication via the platform determined by the hybrid nature of the system: the weekly gatherings at face-to-face sessions made the interactions artificial via the platform. This caused a “diversion" effect ("catachresis" according to Rabardel, 1995) [3]: very little used as an exchange tool (which is one of its main functions), the platform has instead served as a tool to progressive management for achieving group project.

Note also that tasks exchange goes beyond chat sessions especially in exchange forums with the teacher and the one among learners. Another item that evaluates the interaction through the platform comes from the participation in the Forum. Almost 85% (very useful and helpful answers) of students believe that the use of the forum as a tool is useful.

Through the forum, students can ask questions that the teacher or other students can answer and these questions are sometimes mentioned during face-to-face sessions.

This Forum is an opportunity to represent each student and to trigger interactions during training. The forum allows students to submit documents or web links that seem interesting and that could be useful for their peers. Where each of the proposed tasks leads to a discussion in a forum and thus in a public space, it is also about empowering students to allow them be in charge of their training and to get them to discuss issues that were mentioned during the course or exchange notions that would further the development of the face-to-face classes.

A mid-term evaluation of this blended learning experience allowed us to identify some ethical issues at the forum (use of SMS language or the problem of plagiarism) it is mainly during the exchange on a given subject, the student copies extracts from the internet to put them in the forum. What are the reasons for these deviations among distance learners (FOUL platform)? Among other reasons: it is when the teacher is not there. Because of this lack, less "finished" work submitted at the platform will be witnessed.

B. function of the tutor in the hybrid system

The introduction of FOUL platform requires developing new skills, be it the teacher or the student. The hybrid system uses different teaching methods from those in the "face-to-face" training. Kern (2006) [4] points out the importance of training teachers because; "Rethinking the role of teacher means rethinking teacher training."

Two major functions are distinguished for teachers who engage in hybrid training systems: instructional designer and tutor. Regarding the first, Pothier (2003) [5] believes that teachers are the best people for this kind of function as they hold three kinds of expertise needed for this: the content, the likely reactions of learners and the teachers on site aid. The specific nature of online exchanges require their fourth expertise that is attached to the support which it's called technical expertise. Glikman (2002) [6] considers that teachers should have some "technical culture" of the tools used. It is necessary that the teacher knows the implications of the use of the platform: correcting exercises on time (as soon as possible after being submitted in the platform), checking his account on a daily basis if possible (to check e-mail sent by the students), and participation in any forum launched on the platform.

In case of a non active participation of the teacher, students will not see the importance of submitting files if they won't get feedback or correction. Feedback is important for student motivation and good participation on the platform.

In an essay on the introduction of new jobs related to ICT (information and communication technologies), Mangenot (2005, 163) [7] suggests to distinguish four main types of skills: basic technological skills "they will also be a prerequisite for most other functions," those related to the educational support, those related to management and those related to the design of technological resources. So from this experiment, it can be noticed that the skills that a language teacher has to develop are part of the scope of the design of techno-pedagogical tools as well as the online tutoring for teaching French for special purpose.

As part of this experiment, the teacher plays an important role when his functions are of different nature that may be - to use the typology of Rodet (2011) [8] - summarized in the table below:

TABLE III. FIELD OF INTERVENTION OF THE TUTOR / LANGUAGE ADAPTATION (RODET, 2011)

Cognitive	**Metacognitive**	**Motivational**	**Socioaffectifve**
Indicate objectives	Explain the importance of each activity	Encourage participation	Build interest
Work on the disciplinary content	Facilitate planning	Reinforce motivation	Show presence
Find a suitable methodology	Evaluate strategies	Encourage and congratulate	Customize participation
Correct and advise	Assist in self evaluation		Facilitate group collaboration

As for the overall impression on the quality of pedagogical supervision during the test, some students have expressed disinterest in hybrid system and were dissatisfied with the lack of coaching and pedagogical support for students:

"I expected to have my uploaded activities corrected, there are activities which were posted but have not been corrected, so we do not know the result ". (Student SMP).

As mentioned above, it noticed that there is insufficient tutorial help in the platform activities. To analyze these constraints related to tutoring, it was first about tracing interactions between tutors / students to analyze tutorship functions of the four tutors (see table below).

TABLE IV. DESCRIPTION OF THE FREQUENCY OF TUTORSHIP FUNCTIONS FROM THE PLATFORM

Tutorship functions	**Tutor 1**	**Tutor 2**	**Tutor 3**	**Tutor 4**
Pedagogical function	+	+	+	+
Organisational function	+	+	+	+
Socio-motivational function	+	-	+	+
Technical function	-	-	-	+
Assessment	-	-	-	-
Metacognitive function	-	-	-	-

(-): absence of tutoring / (+): presence of tutoring

In terms of frequency, there is order of importance: The pedagogic function (dominant), organizational function and socio-motivational function. And with less importance: The evaluation function (Absence), the technical function and metacognitive function (still very marginal).

TABLE V. ANALYSIS OF TUTORSHIP FUNCTIONS FROM THE PLATFORM FOUL

Tutor interventions (Dominant function)	**Analysis**
Technical function	Almost total absence of technical function because there has been frequent intervention of the person in charge of FOUL platform to solve technical problems.
Organisational function	The tutors present at the beginning of training, objectives, guidelines, etc. The tutor organizes time, facilitates the distribution of tasks, and reminds of due dates, writes summaries..... (4 tutors of responsable for this experiment).
Educational function	The presence of this function in the practice of tutors.
Pedagogical function	The presence of this function in the practice of tutors.
Sociomotivational function	The presence of this function in the practice of tutors: it encourages, gives meaning to the learning goals, creates a friendly environment, creates teamwork spirit, ...
Assessment	No assessment in tracing the tutors' interventions with learners: tutor intervenes in the hybrid system so summative assessment is carried in face-to-face. The trace analysis also revealed that learners participate in the mutual evaluation of the work individually or in groups.

From the table above, certain functions are mobilized enough by all tutors. So there are cardinal functions of tutoring / consensual functions (Denis B, 2003) [9]. These are actually the basic functions: pedagogical, organizational and socio-motivational. The results of this research confirm the presence of these functions to the extent that the later functions remain dominant.

Furthermore, in remote education, it is essential that students have a feedback on their activities to enable them to progress in their learning. Indeed, particular importance was given to the feedback provided on all activities assigned to students. As Rodet stresses (2000, p.46), in remote education, the work of the evaluator is more complex because it "can not limit itself to grading." Thus, the English word "feedback" which indicates that the objective of the evaluater is to give feedback on the work of the learner, Rodet (2000, p.49) [10] defines the term "feedback" as follows: "the feedback comes in response to a work by the learner, (it) offers a guided correction, expresses a value judgment that must be reasoned and argued (and) aims to enable the learner to deepen his knowledge and show him how to do it." But it seemed to us essential to remember is that feedback is "(...) an act of communication which plays a major role in learning" (Rodet, 2000 p.71) [10].

VI. CONCLUSION

In conclusion, it can be deduced that in the course of this research that the transition from classical education to hybrid education is part of a paradigm shift: like any innovation, its acceptance is likely to be slow since it redefines the task and the roles of the teacher and the student. In addition, the success of a hybrid system is conditioned by the development of the tutoring functions of the teacher / tutor. Indeed, the presence of these functions can influence the motivation, creativity and performance of students. Some language teachers may be surprised by the skills they have to implement and by the interventions they are asked to take, such as (the greeting function, technical, conflict resolution within learner teams, etc.). These tutoring functions emphasize the importance of actual training in the role of the tutor. Tutoring is currently conceived as an obligation and not a choice for any university, mainly those longing to further promote the development of e-Learning and the expansion of the hybrid system. Thus, this research has enabled us to formulate results that can be further explored and analyzed in other research projects.

ACKNOWLEDGMENT

I would like to thank to my advisor Mr. Talbi, PhD, for their invaluable guidance and many useful suggestions during my work on this paper. I would also like to express my gratitude to all those who gave me the possibility to complete this paper.

REFERENCES

[1] Gilles Willett, Paradigme, théorie, modèle, schéma : qu'est-ce donc ? [Online] mis en ligne le 26 mars 2012, consulté le 13 mars 2016. URL : http://communicationorganisation.revues.org/1873.

[2] Quintin, J.-J. & Depover, C. & Degache, C , « Le rôle du scénario pédagogique dans l'analyse d'une formation à distance. Analyse d'un scénario pédagogique à partir d'éléments de caractérisation définis. Le cas de la formation Galanet ». EIAH 2005, Montpellier, France.

[3] RABARDEL P, « Les hommes et les technologies. Approche cognitive des instruments contemporains ». 1995, Paris : Armand Colin.

[4] Kern, R, « La communication médiatisée par ordinateur en langues : recherches et applications récentes aux USA », Le Français dans le monde, Recherches et applications, N°40, Les échanges en ligne dans l'apprentissage et la formation, Paris : Clé international, 2006, pp. 17 – 29.

[5] Pothier, M., « Multimédias, dispositifs d'apprentissage et acquisition des langues », 2003, Paris : Ophrys.

[6] Glikman, V, « Des cours par correspondance au E-learning », col. Education et formation, 2002, Paris : PUF.

[7] Mangenot F, « Une formation située de futurs enseignants au multimédia », in Tardieu C. & Pugibet V. Langues et cultures. Les TIC, enseignement et apprentissage, 2005, p. 123-133. Paris, CNDP, Dijon, CRDP de Bourgogne.

[8] Jacques Rodet, Le tuteur à distance et les fonctions d'accompagnement,[Online] mis en ligne Par.jeudi 27 octobre 2011, consulté le 12 mars 2016. URL : http://blogdetad.blogspot.com/2011/10/le-tuteur-distance-et-les-fonctions.html .

[9] Denis B., « Quels rôles et quelle formation pour les tuteurs intervenant dans des dispositifs de formation à distance », in Distance et savoirs, vol. I, n° 1/2003, p. 1-24. DOI : 10.3166/ds.1.19-46

[10] Rodet, J, « La rétroaction, support d'apprentissage ? » 2000, Revue du conseil québécois de la formation à distance, 4.

A Graph Theoretic Approach for Minimizing Storage Space using Bin Packing Heuristics

Debajit Sensarma
Dept. of Computer Science & Engineering
University of Calcutta
Kolkata, India

Samar Sen Sarma
Dept. of Computer Science & Engineering
University of Calcutta
Kolkata, India

***Abstract*—In the age of Big Data the problem of storing huge volume of data in a minimum storage space by utilizing available resources properly is an open problem and an important research aspect in recent days. This problem has a close relationship with the famous classical NP-Hard combinatorial optimization problem namely the "Bin Packing Problem" where bins represent available storage space and the problem is to store the items or data in minimum number of bins. This research work mainly focuses on to find a near optimal solution of the offline one dimensional Bin Packing Problem based on two heuristics by taking the advantages of graph. Additionally, extreme computational results on some benchmark instances are reported and compared with the best known solution and solution produced by the four other well-known bin oriented heuristics. Also some future directions of the proposed work have been depicted.**

Keywords—Bin Packing; Combinatorial Optimization; Graph Theory; Heuristics; Operational Research

I. Introduction

The storage space minimization problem is an open problem of now-a-days as the sizes as well as the dimension of data are increasing day by day. So, there is a need to produce a near optimal solution in less amount of time. To tackle with the problem the author have considered the storage minimization problem as the famous one dimensional Bin Packing Problem where storage space can be represented as bins and the problem is to store the items or data in minimum number of bins. This problem arises in a wide variety of contexts and this popular combinatorial optimization problem has been extensively studied during past few years. The authors [1] called the problem as "The Problem That Wouldn't Go Away". The study of classical one dimensional Bin Packing Problem first begins in the early 1970's [2]. The problem states that, an unlimited number of bins with integer capacity C>0 each, a set of items with their weights, wi, 0< wi ≤ C are given. The goal is to assign each item to one bin, such that total weight of the items in each bin does not exceed the capacity C and the number of bins used for packing all items is minimized. The problem is known to be NP-Hard is strong sense [3]. Thus, in this case the satisfying solution is to design an approximation algorithm which will construct near-optimal packing.

One dimensional Bin Packing Problem has several applications in real world, among them resource and storage space minimization is one facet. Some formulations of real world storage minimization problem using Bin Packing Problem are as follows: i) Placing computer files with specified size into the identical disk with same capacity with constrained that each file must be entirely on one disk [4]. The objective is to minimize the number of disks needed for the set of files. This can be formulated using Bin Packing Problem where items are files, disks are bins and disk capacity is the bin capacity which is fixed. The problem is to minimize the number of bins. ii) Server Consolidation [5] is an approach to the efficient usage of computer server resources in order to reduce the total number of servers or server locations that an organization requires. In this case, existing servers can be treated as items, resource utilizations are item sizes, bins are destination servers and the bin capacity is the utilization threshold of the destination servers. The goal is to minimize the destination servers and maximizing resource utilization. With one resource the problem is same as one dimension Bin Packing Problem. Additionally, with more than one resource (e.g. CPU, disk, memory and computer network) the problem dimension increases. iii) Also the Bin Packing Problem can be used to minimize the cost of storing data (items) in the cloud storage [6]. As buying hard-drive in bulk is much cheaper than buying them individually, the goal of solving the problem becomes minimizing the hard-drives (bins) to store the data (items). Besides this, there are other storage minimization problems where Bin Packing has a major role, but are not discussed in this paper.

Not only Bin Packing Problem but also graph theory has vast real world applications. Graph algorithm provides unified solution approach to many classical and modern application areas by taking graph as an omnipotent mathematical tool. In view of storage minimization problem, there exists various graph compression mechanism which can be used to store data compactly [7].

This paper mainly focused on the solution of one dimensional Bin packing problem in polynomial time, and for this an algorithm depending on two offline bin oriented heuristics has been proposed taking the advantages of graph theory. Firstly, a vertex weighted graph is constructed from the set of item weights where for each item weights one vertex is created. Then, the first heuristic chooses the subset of vertices according to the maximum total weight criteria and the second one is based on maximum average weight criteria, which ultimately produces the minimal clique partition of the graph with each clique having weight not exceeding the capacity of each bin. The total number of partition gives the total number of bins. The algorithm runs in polynomial time.

Most of the existing algorithms not completely based on graph algorithm rather hybridization of graph algorithms but this work is completely based on a graph algorithm to find minimum clique Partition with weight constraint and can compete with existing algorithms. Also it can open a new direction for solving multi-dimensional Bin Packing Problem. The detailed description of the algorithm can be found in subsequent sections.

The article is organized as follows: section II contains some preliminary concepts related to the work. Some existing work to tackle Bin Packing problem with graph is described in section III. Section IV gives the detailed description of proposed algorithm. Section V contains computational results. Finally, section VI concludes the article giving some future scopes in section VII.

II. Preliminaries

This section contains some preliminary concepts related to the topic, taken from [8, 9, 10, 11].

Definition 2.1: A **Graph** G is a triple consisting of a vertex set V (G), an edge set E(G), and a relation that associates with each edge two vertices (not necessarily distinct) called its endpoints.

Definition 2.2: vertex weighted graph is a graph where each vertex has been assigned a positive weight.

Definition 2.3: A **Null Graph** is a graph whose edge set is null.

Definition 2.4: A **Clique** in a graph G is a set of pairwise adjacent vertices.

Definition 2.5: A vertex x of a graph G is **Simplical Vertex** if its adjacency set Adj(x) induces a complete subgraph of G, i.e. Adj(x) is a clique (not necessarily maximal).

Definition 2.6: An ordering δ = [v_1, v_2, …, v_n] where n is the number of vertices of an undirected graph G is **Perfect Elimination Ordering** iff each v_i is a simplical vertex of the induced sub graph $G_{\{vi,\dots vn\}}$.

Definition 2.7: Chordal Graph is a simple graph in which every cycle of length four and greater has a cycle chord.

Definition 2.8: Given a vertex weighted graph G = (V, E; W}, having weight of vertices w_1, w_2,…, $w_{|V|}$ respectively and a bound C′, the **Minimum Clique Partition with Constrained Weight (MCPCW)** problem [12] is to find a partition of these |v| vertices into smallest number of cliques such that each clique has its weight not beyond C′.

Theorem 2.9: Minimum Clique Partition with Constrained Weight (MCPCW) problem is NP-Hard.

Proof. The proof is done by transforming an instance of 3-Partition problem to an instance of **MCPCW**.

Consider an instance P of 3-Partition problem: Given the set S = {a_1, a_2, …, a_{3k}}of 3k integers satisfying C′/4 < a_j < C′/2 for each 1 ≤ j ≤ 3k and $\sum_{j=1}^{3k} a_j$ = kC′. The problem asks whether S can be partitioned into k subsets s_1, s_2,…, s_k, such that for each i= 1, 2,…, k, s_i contains exactly three elements of S and $\sum_{a \in s_i} a$ = C′.

Now we will construct a polynomial time reduction Q for P of the 3-partition problem to an instance Q(P) of the **MCPCW** problem i.e. a vertex weighted graph with weight of each vertices w_1, w_2, …, w_{3k} respectively where w_i = a_i for each i= 1,… 3k and the bound C′= ($\sum_{j=1}^{3k} w_j$)/ k.

We now prove the claim that there exists a feasible solution to an instance P of the 3-partition problem iff instance Q(P) of the **MCPCW** problem has its optimal solution. So, the feasible partition of the instance Q(P) can be constructed in the following way: for each s_i = { a_{i_1} , a_{i_2} ,…, a_{i_l} }, select the clique c_1 = { w_{i_1} , w_{i_2} ,…, w_{i_l} } and then obtains a partition of these 3k weights of |V| vertices into k cliques having weight exactly C′. Conversely, if the instance Q(P) of the problem **MCPCW** has an optimal clique partition { c_1, c_2, …, c_k} with the smallest integer k having $\sum_{w \in c_i} w \leq$ C′ for each i=1,…,k. By the facts, $\sum_{j=1}^{3k} a_j$ = kC′ and C′/4 < a_j < C′/2 for each 1 ≤ j ≤ 3k, we obtain $\sum_{w \in c_i} w$ = C′ (as w_j = a_j for each j= 1,… 3k) for each j= 1, 2,…, k and the clique c_j contains exactly three elements from S, i.e. s_j = { a_{j_1} , a_{j_2} , a_{j_3} } and $\sum_{a \in c_j} a$ = C′ j= 1,… k. So, the instance p of the 3-partition problem has the partition s_1, s_2,…, s_k.

Definition 2.10: A **bin oriented heuristic** for Bin Packing Problem constructs solution bin by bin i.e. while unpacked items remain it is packed with the maximal subset of unpacked items, e.g. First Fit Decreasing (FFD), Best-Two-Fit (B2F), Minimum Bin Slack (MBS), MBS′ etc. [13, 14].

Definition 2.11: Offline algorithms have all the items available before the packing starts, e.g. First Fit Decreasing (FFD) [4].

III. Related Works

This section consists of some related works to solve one dimensional Bin Packing Problem based on graphs. Firstly, in [15] the authors consider time constrained scheduling problem. For a set of jobs J with execution time t(j)∈(0, 1] and an undirected graph (the conflict graph) G =(J, E), they consider to find schedule of the jobs that are adjacent and they are assigned different machines (bins) with total execution time of each machine at most 1. The objective is to assign all jobs into minimum number of machines maintaining the time constraint. To tackle the problem, they have proposed six different algorithms based on different principles. The first three algorithms are the modification of classical NF, FF, FFD algorithms. Next algorithm depends on optimal coloring algorithm which finds a minimum partition of the item set into independent sets which is equal to the chromatic number of G and applies one of the NF, FF and FFD packing to each independent set. Fifth and sixth algorithm is same like above but the main difference is fifth one is based on pre-coloring method and sixth one is based on general coloring method that works for co-graph and k-trees. Next, the authors of [16] consider the problem namely Bin Packing with Conflict (**BPC**) using conflict graph and it's online, offline versions. They mainly improve the upper bounds of BPC on perfect graphs, interval graph and bipartite graphs. Most of the recent results follow from the adaptation of weighting systems to enable analysis of algorithms for BPC and new algorithms which carefully remove small sub-graph of items causing problematic instances. In next work [17] authors considered a restricted problem called Bin Packing with Clique Graph Conflicts. They have designed a polynomial time approximation algorithm for constant item size analyzing its performance in the more general case of bounded item sizes. In [18] authors investigated the following problem: the items to be packed are structured as the leaves of a tree and it is called as Structured Bin Packing Problem. The objective is to pack the items in the same bin whose lowest common ancestor has low height. Next, authors of [19] have proposed a problem to pack a graph G with lower and upper bound on its edges and weights on its vertices into a host graph I and called the problem as Graph Bin Packing Problem. The vertices of G are items to be packed and vertices of I are bins. The host vertex can accommodate at most L weight in total and if two items are adjacent in G, then the distance of their host vertices in I must be between the lower and upper bounds on the edge joining the two items.

Most of the above algorithms not completely based on graph algorithm rather hybridization of graph algorithms and exiting heuristics for solving Bin Packing Problem. Our work is simple and purely based on a graph algorithm namely finding minimum clique Partition with weight constraint and can compete with existing algorithms. Also it can open a new direction for solving multi-dimensional Bin Packing Problem.

IV. The Proposed Algorithm

Let, W= {w_1, w_2, …,w_n} be the given sequence of weights of the items. The items are numbered 1 through n, from the left to right of the list, labeling their positions in W, i.e. w_1 is the weight of first item in W, w_2 is the weight of second item in W and so on.

In this section an algorithms based on two bin oriented heuristics has been formulated based on graph to cope with the one dimensional Bin Packing Problem.

In this algorithm firstly items are sorted in non decreasing order with respect to their weight. Next, a vertex weighted graph is constructed from the sequence of items. Here, for each item a weighted vertex are introduced. Hence, firstly the graph consists of 'n' isolated vertices {v_1,…,v_n} with their weight {w_1, w_2, …,w_n} respectively. Now, for introducing edges to the graph, the following procedure is being followed. For any pair of items with weight w_i and w_j that are in the position i and j, respectively in W, an edge is introduced between the corresponding vertices of w_i and w_j, only if (i-j)(w_i-(C-w_j)) ≥ 0, where C is the capacity of each bin. In other words, an edge is introduced between the corresponding vertices in the graph if they satisfy the condition $w_i + w_j \leq C$. This is explained with an example below.

Example 4.1: Suppose, W = {8, 11, 10, 4, 7, 9, 3}, C = 15.

After sorting the sequence is W′ = {11, 10, 9, 8, 7, 4, 3}.

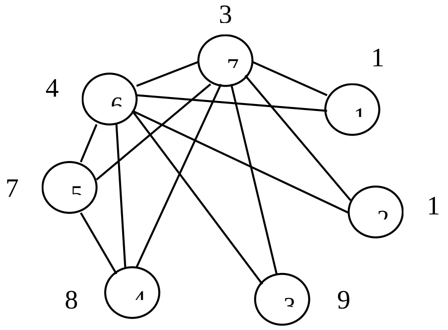


Fig. 1. Vertex weighted graph for the sequence W = {8, 11, 10, 4, 7, 9, 3} and bin capacity C=15. Vertex 1 has weight 11, vertex 2 has weight 10 and so on. An edge {vi, vj} indicates that wi + wj ≤ C.

Lemma 4.2: The Graph produced from the sequence W′ after sorting the sequence W in non increasing order (i.e. $w_1 \geq w_2 \geq \ldots \geq w_n$), has a Perfect Elimination Ordering.

Suppose, $W_i \geq W_j \geq W_k$ and vertex i and j are connected. The following equations are satisfied.

$W_i + W_j \leq C$ … (1)
$W_i + W_k \leq C$ … (2)

If the ordering is the perfect elimination ordering then, vertex j and k will also be connected and $W_j + W_k \leq C$.

Adding (1) and (2)

$2W_i + (W_j + W_k) \leq 2C$

Or, $W_j + W_k \leq 2(C-W_i)$… (3)

As, $W_j \leq W_i$

Putting $W_j = W_i$ we get from (1)

$2W_i \leq C$

From, (3) we get,

$W_j + W_k \leq C$

This condition is applicable for the whole ordering. So, the ordering is the perfect elimination ordering.

Claim 4.5: There exists a feasible solution to an instance I of one dimensional Bin Packing Problem if and

only if the instance τ (I) of the MCPCW problem for Chordal Graph has its optimal solution with value k.

For any feasible solution of an instance I of Bin Packing Problem, the set of items B is partitioned into k bins $\{B_1, B_2, \ldots, B_k\}$, $\forall$ i=1, …, k, such that each B_i contains items of B and $\sum_{a \in B_i} a \leq C$ (C=capacity of each bin). Then a feasible partition of the instance τ (I) of the MCPCW problem can be constructed in the following way: for each $B_i = \{ a_{i_1}, a_{i_2}, \ldots, a_{i_l} \}$, select the clique $C_i = \{ w_{i_1}, w_{i_2}, \ldots, w_{i_l} \}$ having total weight of the vertices not exceeding C. Likewise obtain the partition of total items into k cliques each having total weight not exceeding C.

Conversely, if the instance τ (I) of the MCPBW problem has optimal clique partition $\{C_1, C_2, \ldots, C_k\}$, $\forall$ i=1, …, k, with smallest integer k and having $\sum_{a \in c_i} a \leq C$. Then each clique contains items from the set B, i.e. $B_j = \{ a_{i_1}, a_{i_2}, \ldots, a_{i_l} \}$ and $\sum_{a \in c_j} a \leq C$, $\forall$ i=1, …, k. So, the instance I of Bin Packing Problem has the partition $\{B_1, B_2, \ldots, B_k\}$.

Algorithm 4.1: Counting Bins

```
Input: List of vertices (n) with their weights {$w_1, w_2, \ldots, w_n$}, Capacity (C).
Output: Number of Bins (B).
Begin
Step 1: If (n! = 0) then go to step 2 else goto step 7.
Step 2: Sort the vertices according to non-increasing order of their weight.
Step 3: Call Algorithm 4.1.1.
Step 4: Call Algorithm 4.1.2.
Step 5: Assign clique partition number (obtained from step 4) of vertex weighted
graph (G) (produced by step 3) with total weight of each clique $\leq$ C to B (i.e. B $\longleftarrow$ CC (Clique Count)).
Step 6: Print B.
Step 7: End.
```

Algorithm 4.1.1: Construct_ Graph

```
Input: List of vertices with weights {$w_1, w_2, \ldots, w_n$}, Capacity (C).
Output: Vertex weighted Graph (G).
Begin
Step 1: Set i $\longleftarrow$ 1, j $\longleftarrow$ 1.
Step 2: If ( i $\leq$ n) then goto step 3 else goto step 9.
Step 3: If (j $\leq$ n) then goto step 4 else goto step 8.
Step 4: If ( i $\neq$ j ) goto step 5 else goto step 7.
Step 5: If ($w_i$ + $w_j \leq$ C) then goto step 6 else goto step 7.
Step 6: Connect item i and j.
Step 7: Set j $\longleftarrow$ j+1, goto step 3.
Step 8: Set i $\longleftarrow$ i+1, goto step 2.
Step 9: End
```

Algorithm 4.1.2: Minimum Clique Partition with Constrained Weight (MCPCW)

```
Input: Adjacency List of the Vertex weighted Graph (G), Capacity (C).
Output: Clique Count (CC).
Begin
Step 1: CC $\longleftarrow$ 0;
Step 2: If (n!= 0) then goto step 3 else goto step 7;
Step 3: i $\longleftarrow$ 1;
Step 4: If vertex i has zero or one neighbor, then delete the vertex along with its
neighbor (if any) from the Graph (G), CC $\longleftarrow$ CC + 1 and goto step 2 else goto step 5;
Step 5: Select subset of vertices consisting of vertex i and its neighbor vertices based on Selection criteria 1 or Selection Criteria 2.
Step 6: Delete the subset produced from step 4, CC $\longleftarrow$ CC + 1, goto step 2;
Step 7: End
```

As the subsets are the cliques, so algorithm 4.1.2 returns the number of clique partition with each partition weight not exceeding the capacity. The critical part of the algorithm 4.1 is step 4 of the algorithm 4.1.2 where subset of the vertices consisting of the current vertex and its neighbors has to be selected. Here, we have adopted two heuristics for selection of the subset. The selection criteria are depicted below:

A. *Selection Criteria 1 (A1):*

This criterion selects the subset of the current vertex along with its neighbor vertices which gives maximum total weight not exceeding the capacity (**C**).

B. *Selection Criteria 2 (A2):*

This criterion selects the subset of the current vertex along with its neighbor vertices which gives maximum average weight not exceeding the capacity (**C**). Here, firstly the total average weight (**T_a**) of the vertex set is calculated. Suppose, average weight of current subset is **C_a** and average weight of its previous subset is **P_a**, then if $C_a \geq P_a$ or $C_a \geq T_a$ and also total sum of current subset is greater than the previous one, current subset is selected as the final subset, otherwise previous subset is selected as the final subset and this process continues for all possible subsets.

Theorem 4.3: The graph G formulated by the Algorithm 4.1, is a Chordal Graph.

Proof. Let, there is a chordless cycle $v_1, v_2, \ldots, v_l$, with $l \geq 4$ in G. According to lemma 4.2, the graph G has perfect vertex elimination ordering. Suppose, v_i is the vertex in the cycle that occurs first in the perfect elimination ordering and v_{i+1}, v_{i+2} are neighbors of v_i occur later in the ordering. So, there must be an edge between v_{i+1} and v_{i+2}. But this contradicts the assumption that the cycle is chordless. So, the graph G is a Chordal Graph.

Claim 4.4: Any induced subgraph of the graph G produced by the Algorithm 4.1, is Chordal.

As the graph G produced by the Algorithm 4.1 is Chordal and any induced subgraph of a Chordal Graph is Chordal [20], so the above claim is also true for the graph produced by the Algorithm 4.1.

Lemma 4.6: Minimum Clique Partition Problem with Constrained Weight (MCPCW) for the Chordal Graph can be solved in O(|V| +|E|) time where V is the vertex set and E is the edge set.

According to lemma 4.2, the ordering $w_1 \geq w_2 \geq \ldots \geq w_n$ is a perfect vertex elimination ordering. Suppose, processing starts with vertex v_1 with weight w_1. It is added to the first partition. Next the adjacent vertices of v_1 are checked and the vertices are added along with v_1 to the partition with total weight ≤ C, based on one of the two above selection criteria. If the first partition is $\{v_1, v_{2, \ldots,} v_k\}$, then after deletion of the vertices in the partition Algorithm 4.1 continues the execution with the remaining graph G´, which is also a Chordal Graph according to the lemma 4.4. The execution continues until vertex set is empty. In each iteration, Algorithm 4.1 checks the vertex and its neighbors. So, the overall complexity of the implementation is $O(|v|) + O\sum_{v\in V}(|\,\mathrm{Adj}(v)\,|)$, which is roughly equivalent to O (|V| + |E|).

Theorem 4.7: Number of Bins Produced by the Algorithm 4.1 is K ≤ 3/2 OPT +1 and time complexity is O (|V|^2).

Proof. Assume, partition of the ordered list of vertex weights has to be done where the weights $\{\mathbf{w_1, w_2, \ldots, w_n}\}$ are distributed in the following sets:

X = {w_i | w_i > 2L/3} {L=capacity of each Bin}

Y = {w_i | L/2 <w_i ≤ 2L/3}

T = {w_i | L/3 <w_i ≤ L/2}

Z = {w_i | w_i ≤ L/3}

Case 1: There is one Clique Partition with all vertices from set Z.

1) In this case all partitions except the last one have used more than 2C/3 of the total capacities. Otherwise an item from set Z can put into them.

2) It has to be the last partition.

Suppose, required number of bins = K.

$\therefore 2L(K-1)/3 + (C_K) \leq \sum_{i=1}^{n} w_i$ [C_K = total weight of partition K]

$\Rightarrow 2(K-1)/3 + (C_K)/L \leq \left\lceil \sum_{i=1}^{n} w_i/L \right\rceil \leq$ OPT [$\because$ OPT= $\left\lceil \sum_{i=1}^{n} w_i/L \right\rceil$]

$\Rightarrow K \leq 3/2$ OPT + 1 -3/2. C_K/L [Clique Partition contains one vertex with weight= L/3]

$\Rightarrow K \leq 3/2$ OPT + 1 -1/2

$\Rightarrow K \leq 3/2$ OPT + 1

Case 2: There is no Clique Partition with all vertices from Z.

In this case all vertices from set Z can be thrown out without changing total number of partitions and below cases arise.

1) No partition has more than 2 items.

2) Any partition with one vertex from X cannot accommodate any other vertices.

3) Any partition with one vertex from Y can accommodate only another vertex from T.

4) Any Partition with one vertex from T can accommodate either one vertex from Y or one vertex from T but not both.

From the conclusion above we know that know that our algorithm will put at most 2 vertices in a bin. So, it put each vertex in a partition with maximum total weight (criteria 1) and maximum average weight (criteria 2). So, in this case the solution of proposed algorithm is optimal.

For the second part, it can be seen from the algorithm that for V number of vertices algorithm 4.1.1 construct the graph in O(|V|^2) time and from Lemma 3.5 it can be concluded that algorithm 4.1.2 requires O(|V| + |E|) time. As time complexity of algorithm 4.1.1 dominates time complexity of algorithm 4.1.2; total time required by the algorithm 4.1 is O(|V|^2).

C. Illustration of Algorithm 4.1 (Counting Bins) with examples

Suppose, set W is the set of vertex weights organized in non-increasing order of their sizes and $w_i \in Z^+ \ \forall$ i=1, 2 …, n. W= $\{w_1 \geq w_2 \geq \ldots \geq w_s > C/2 > w_{s+1} \geq w_{s+2} \geq \ldots \geq w_n\}$ and capacity=C. The optimal number of bins is calculated as $\left\lceil \sum_{i=1}^{n} w_i/C \right\rceil$.

Case 1: Input sequence: $y_1 \geq y_2 \geq \ldots \geq y_s > C/2$

The graph can be viewed as a null graph and in that case the number of cliques is the cardinality of the vertex weight set which is |W|.

Example 4.8: W = {15, 14, 12, 11, 10, 9, 8} and C=15.Optimal No. Bins= 6.

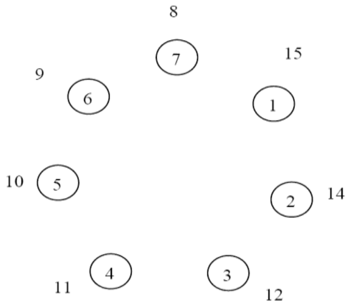


Fig. 2. Vertex weighted graph for case 1

So, with selection criteria 1 and selection criteria 2, the number of Bins=7.

Case2: Input sequence: $w_1 \geq w_2 \geq \ldots \geq w_s > C/2 > w_{s+1} \geq w_{s+2} \geq \ldots \geq w_n$

This graph is a Chordal graph. Number of bins can be found by finding minimum clique partition with total weight of each clique not exceeding C.

Example 4.9: W = {11, 10, 9, 8, 7, 4, 3} and C=15. Optimal no. of Bins=4

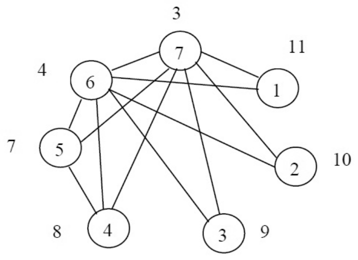
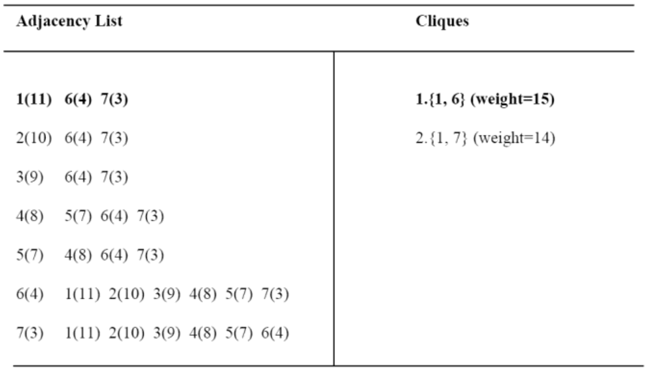


Fig. 3. Vertex weighted graph for case 2

Selection Criteria 1:



Adjacency List	Cliques
1(11) 6(4) 7(3)	**1.{1, 6} (weight=15)**
2(10) 6(4) 7(3)	2.{1, 7} (weight=14)
3(9) 6(4) 7(3)	
4(8) 5(7) 6(4) 7(3)	
5(7) 4(8) 6(4) 7(3)	
6(4) 1(11) 2(10) 3(9) 4(8) 5(7) 7(3)	
7(3) 1(11) 2(10) 3(9) 4(8) 5(7) 6(4)	

Fig. 4. Adjacency list of the graph in Figure.3 and possible cliques

Clique (C_1) = (11, 4).

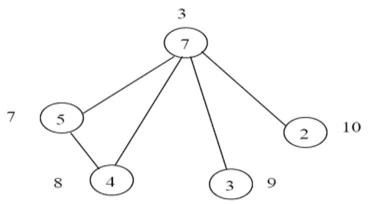
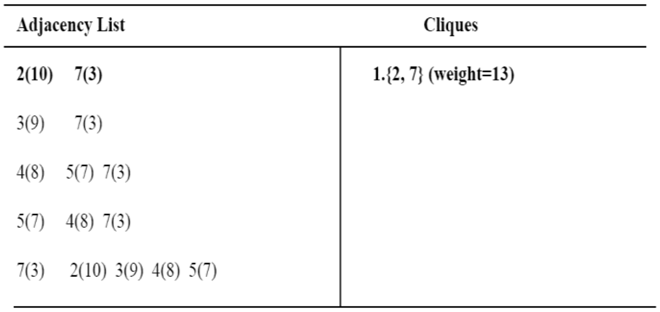


Fig. 5. Vertex weighted graph after deletion of vertex labeled 1 and 6



Adjacency List	Cliques
2(10) 7(3)	**1.{2, 7} (weight=13)**
3(9) 7(3)	
4(8) 5(7) 7(3)	
5(7) 4(8) 7(3)	
7(3) 2(10) 3(9) 4(8) 5(7)	

Fig. 6. Adjacency list of the graph in Figure.5 and possible cliques

Clique (C_2) = (10, 3).

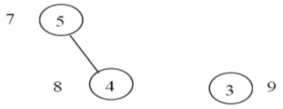
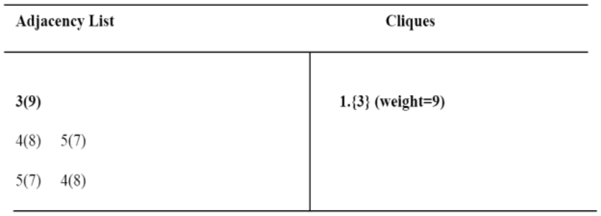


Fig. 7. Vertex weighted graph after deletion of vertex labeled 2 and 7



Adjacency List	Cliques
3(9)	**1.{3} (weight=9)**
4(8) 5(7)	
5(7) 4(8)	

Fig. 8. Adjacency list of the graph in Figure.7 and possible clique

Clique (C_3) = (9).

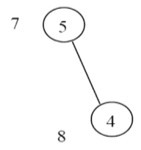


Fig. 9. Vertex weighted graph after deletion of vertex labeled 3

Adjacency List	Cliques
4(8) 5(7)	**1.{4, 5} (weight=15)**
5(7) 4(8)	

Fig. 10. Adjacency list of the graph in Figure.9 and possible clique

Clique (C_4) = (8, 7).

So, total number of bins C_1, C_2, C_3, C_4 = 4.

Selection Criteria 2:

Total average weight of the vertices T_a = **(**11+ 10+ 9+ 8+ 7+ 4+ 3)/7 =7.43.

Adjacency List	Cliques
1(11) 6(4) 7(3)	**1.{1, 6} (Average weight = (11+4)/2 = 7.5 > T_a and weight=15)**
2(10) 6(4) 7(3)	2.{1, 7} (Average weight= (11+3)/2 = 7 <T_a and weight=14)
3(9) 6(4) 7(3)	
4(8) 5(7) 6(4) 7(3)	
5(7) 4(8) 6(4) 7(3)	
6(4) 1(11) 2(10) 3(9) 4(8) 5(7) 7(3)	
7(3) 1(11) 2(10) 3(9) 4(8) 5(7) 6(4)	

Fig. 11. Adjacency list of the graph in Figure.3 and possible cliques

Clique (C_1) = (11, 4).

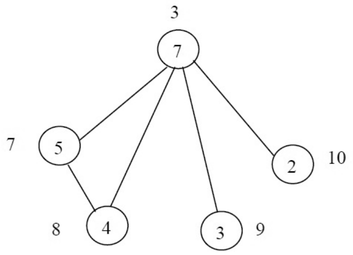


Fig. 12. Vertex weighted graph after deletion of vertex labeled 1 and 6

Adjacency List	Cliques
2(10) 7(3)	**1.{2, 7} (Average weight=(10+3)/2 =6.5 < T_a and weight=13)**
3(9) 7(3)	
4(8) 5(7) 7(3)	
5(7) 4(8) 7(3)	
7(3) 2(10) 3(9) 4(8) 5(7)	

Fig. 13. Adjacency list of the graph in Figure.12 and possible clique

Clique (C_2) = (10, 3).

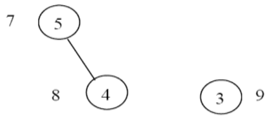


Fig. 14. Vertex weighted graph after deletion of vertex labeled 2 and 7

Adjacency List	Cliques
3(9)	**1.{3} (Average weight 9 > T_a and weight=9)**
4(8) 5(7)	
5(7) 4(8)	

Fig. 15. Adjacency list of the graph in Figure.14 and possible clique

Clique (C_3) = (9).

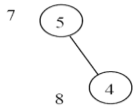
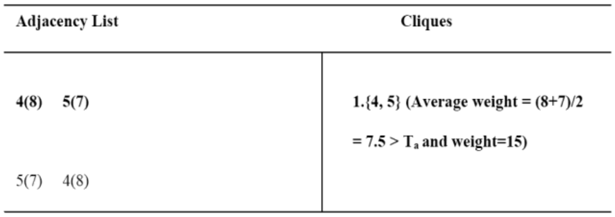


Fig. 16. Vertex weighted graph after deletion of vertex labeled 3



Adjacency List	Cliques
4(8) 5(7)	1.{4, 5} (Average weight = (8+7)/2 = 7.5 > T_a and weight=15)
5(7) 4(8)	

Fig. 17. Adjacency list of the graph in Figure.16 and possible clique

Clique (C_4) = (8, 7).

Here also total number of bins= 4.

Case 3: Input sequence: $C/2 > x_{s+1} \geq x_{s+2} \geq \ldots \geq x_n$

In this case the graph can be viewed as a clique, which is also a Chordal graph.

Here also the number of bin is the number of minimal clique partition with total weight of each clique ≤ C.

Example 4.10: W= {10, 9, 9, 9, 8, 8, 7, 7} and C=24. Optimal number of Bins = 3.

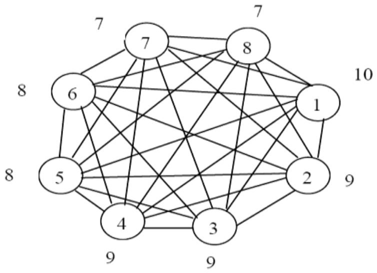


Fig. 18. Vertex weighted graph for case 3

Applying selection criteria 1 (A1) we get the number of bins= 4, i.e. clique (C_1) = (10, 7, 7), clique (C_2) = (9, 9), clique (C_3) = (9, 8) and clique (C_4) = (8), but with selection criteria 2 (A2), the number of bins= 3, i.e. clique (C_1) = (10, 9), clique (C_2) = (9, 8, 7), clique (C_3) = (9, 8, 7) which is improved than former.

V. Computational Results

The proposed algorithm was coded in C, compiled using Borland C++ 5.0 compiler in Win32 mode and in Intel® Atom™ 1.60 Hz Processor with 1.0 GB DDR2 RAM.

The algorithms were tested on six classes of benchmark problem instances, all of which can be downloaded from the web page of EURO Special Interest Group on Cutting and Packing (ESICUP) (http://paginas.fe.up.pt/~esicup/). The propose algorithm with heuristic criteria 1 is named as A1 and with heuristic criteria 2 is named as A2.

The first two, the u class and t class, were developed by [21] and named instance ‘a’ in table I. The u class has item weights drawn from an integer uniform distribution on (20, 100) and bin capacity c= 150. There are four sets in this class, namely u_120, u_250, u_500 and u_1000; each consisting of 20 instances with n= 120, 250, 500 and 1000 items, respectively. The t class has item weights drawn from a uniform distribution on (25, 50) and c= 100. Item weights in this class are real numbers. There are also four sets in this class, namely t_60, t_120, t_249 and t_501; each consisting of 20 instances with n= 60, 120, 249 and 501 items, respectively. The t class is considered difficult, because in an optimal solution of each instance, each bin contains 3 items with zero slack (hence the name ‘triplets class’). All problem instances in both the u and t classes have been solved to optimality with the exact algorithm of [22]. It can be seen from table I that, proposed A1 and A2 finds the solution better than FFD heuristic and A1 is giving better solution than A2.

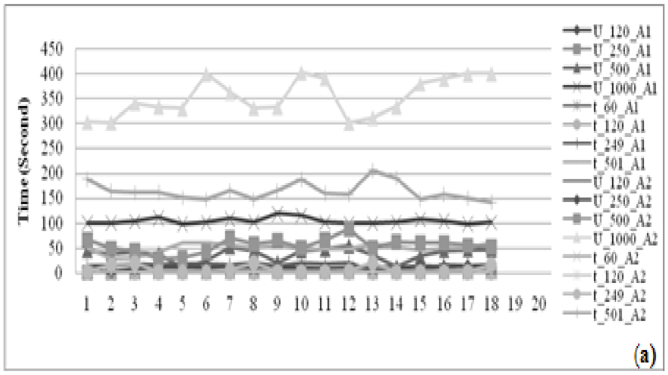
A third class of benchmark problem instances, developed by [23], contains two sets, was_1 and was_2 and named instance ‘b’ in table I. Each set has 100 instances with c= 1000 and item weights from (150,200). Was_1 has n= 100 items in each instance, while was_2 has n= 120 items. For all instances in this class, optimal solutions are known. Solution produced by the proposed A1 and A2 are better than FFD heuristic but A2 has better solution than all other heuristics in table I.

A fourth class of benchmark problem instances, developed by [24], is called gau_1 and contains 17 problem instances with c= 10,000 and various values of n and item weights. It is named instance ‘c’ in table I. For all instances the optimality gap is one bin. Solutions produced by proposed A1 and A2 are same and are better than FFD and B2F heuristics.

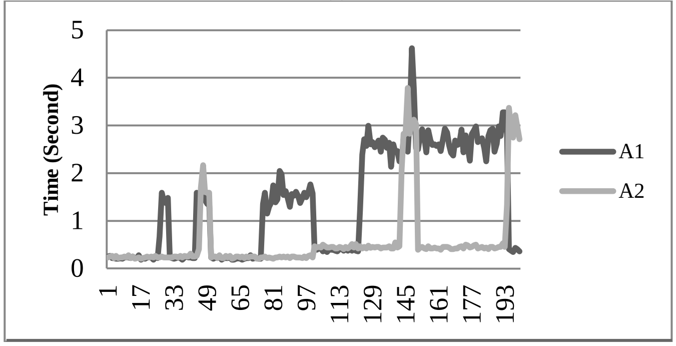
Next, the test on the data set of difficult problem instances has been performed, called hard28, used for example by [25] and named instance ‘d’ in table I. This set has 28 instances with n∈{160,180, 200}, c= 1000, and items weights drawn from (1,800). The simplest heuristic, FFD, finds optimal solutions for five instances and solutions worse than optimal by one bin for all the remaining instances. None of the other heuristics including A1, is able to improve these solutions. In fact, B2F, MBS, MBS′, A1 find worse solutions for some instances. But proposed A2 finds the same solutions as FFD.

A sixth class of benchmark problem instances, developed by [26], consists of set_1, set_2, and set_3 and named instance ‘e_1’, ‘e_2’, ‘e_3’ in table I respectively. Set_1 has 720 instances with c= 100, 120, 150, n= 50, 100, 200, 500, and item weights drawn from an integer uniform distribution on (1,100), (20, 100), and (30, 100). Set_2 has 480 instances with c= 1000, n= 50, 100, 200, 500 and item weights such that each bin has on average 3 to 9 items. Set_3 has 10 instances with c= 100,000, n= 200, and item weights drawn from a uniform distribution on (20,000, 35,000). Optimal solutions for 1184 instances in this class have been found in [26]. For the remaining 26 instances, optimal solutions were found by [27]. From Table I it can be seen that solution produced by proposed A2 is better than other heuristics.

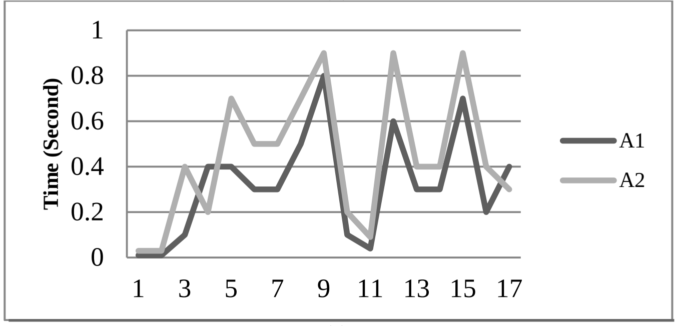
Additionally, in Figure.19 time comparison between two proposed heuristics A1 and A2 for above instances are shown.



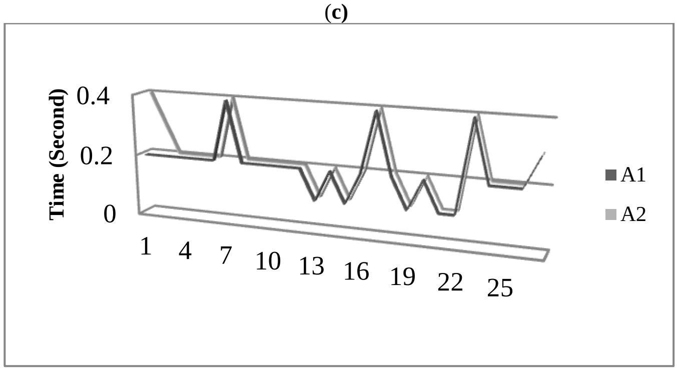
(a)



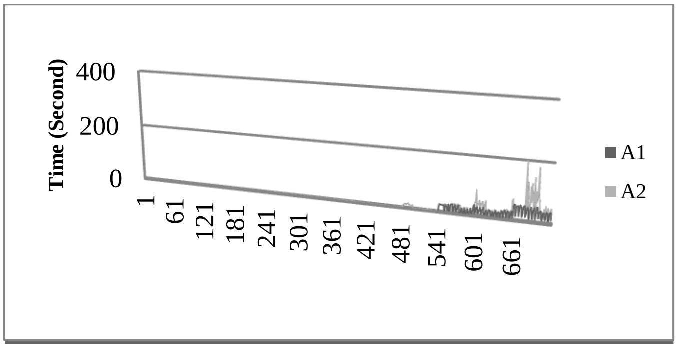
(b)



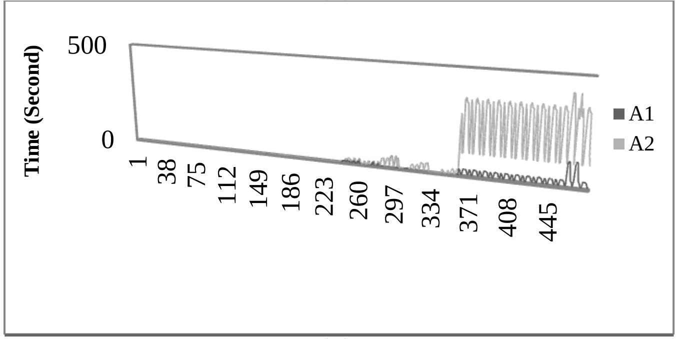
(c)



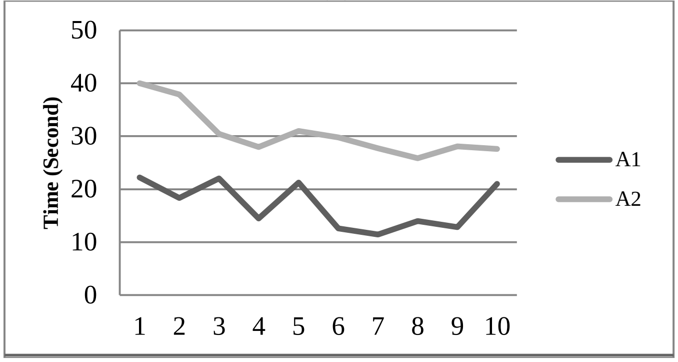
(d)



(e₁)



(e₂)



(e₃)

Fig. 19. Time Comparison between two proposed heuristics A1 and A2 for instances a, b, c, d, e1, e2, e3 respectively. X-axes represent time and Y-axes represent number of instances

VI. Conclusions

It is the very beginning stage of the solution of the problem to store large amount of data in a minimum storage space. In this paper, mainly one dimensional Bin Packing Problem has been treated by a graph algorithm. The proposed algorithm is based on two heuristics; one is depending on maximum total weight criteria of the vertices not exceeding the bin capacity and second is based on maximum average weight criteria of the vertices also not exceeding the capacity of the bin. The algorithm takes polynomial time and finds near-optimal solution which is shown in computational results. It can also be seen that, no algorithm is better for all the instances, the algorithm with second criteria (A2) outperforms other heuristics for one benchmark instance and in other cases, solution with two heuristics (A1 and A2) deviates small from the best known solutions and solutions produced by the four heuristics.

VII. Future Work

Several further research scopes can be outlined as follows. Firstly, along with the volume of the data its dimension is also increasing. To tackle this problem good and efficient algorithms are needed for storing high dimensional data. In this case, multidimensional network graph concept can be used or vertex weighted graph for each dimension can be created and the proposed algorithm can be applied to the graph produced from the intersection of graphs of each dimension. But it needs further investigations. Secondly, online partition problem or semi-online partition problem [28] can be applied to generate cliques with constrained weight of the vertex weighted graph. This concept can be used to tackle online or semi-online Bin Packing Problem but needs further detailed study. Third, running time of the algorithm can be improved from O ($|V|^2$) to O (|V|log|V|) by using the appropriate data structure namely red black tree [29] which is under the investigation of the author and last but not the least, though this paper contains a study of the bound of maximum number of bins needed by the proposed algorithm, the proof of the upper bound involves an extremely detailed case analysis which can be investigated in future.

Acknowledgements

The authors would like to thank University Of Calcutta, West Bengal, India, Department of Science & Technology (DST- under Grant [**No. DST/ INSPIRE Fellowship/2013/943**]), New Delhi, for financial support and the reviewers for their constructive and helpful comments and specially the Computer without which no work was possible.

References

[1] G. Michael R., and D. S. Johnson, “Approximation algorithms for bin packing problems: A survey,” Analysis and design of algorithms in combinatorial optimization 266 (1981): 147-172.

[2] G. Michael R., R. L. Graham, and J. D. Ullman, “Worst-case analysis of memory allocation algorithms,” In Proceedings of the fourth annual ACM symposium on Theory of computing, pp. 143-150. ACM, 1972.

[3] G. Michael R., and D. S. Johnson, Computers and intractability. Vol. 29. New York: wh freeman, 2002.

[4] T. F. Gonzalez., ed., Handbook of approximation algorithms and metaheuristics. CRC Press, 2007.

[5] R. Gupta, S. K. Bose, S. Sundarrajan, M. Chebiyam, and A. Chakrabarti, “A two stage heuristic algorithm for solving the server consolidation problem with item-item and bin item incompatibility constraints,” In Services Computing, 2008. SCC'08. IEEE International Conference on, vol. 2, pp. 39-46. IEEE, 2008.

[6] D. Bein, W. Bein, and S. Venigella, “Cloud storage and online bin packing,” In Intelligent Distributed Computing V, pp. 63-68. Springer Berlin Heidelberg, 2011.

[7] D. Sensarma, and S. S. Sarma, “A Unified Framework for Security and Storage of Information,” Internationa Journal of Advance Engineering and Research Development, Vol.2, No.1, 2015.

[8] D. B. West, Introduction to graph theory, Vol. 2. Upper Saddle River: Prentice hall, 2001.

[9] N. Deo, Graph theory with applications to engineering and computer science. (1994).

[10] M. C. Golumbic, Algorithmic graph theory and perfect graphs, Vol. 57, Elsevier, 2004.

[11] A. Shapira, R. Yuster, and U. Zwick., “All-pairs bottleneck paths in vertex weighted graphs,” In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 978-985. Society for Industrial and Applied Mathematics, 2007.

[12] J. Li, M. Chen, J. Li, and W. Li., “Minimum clique partition problem with constrained weight for interval graphs,” In Computing and Combinatorics, pp. 459-468. Springer Berlin Heidelberg, 2006.

[13] S. K. Basu, Design methods and analysis of algorithms, PHI Learning Pvt. Ltd., 2013.

[14] K. Fleszar, and C. Charalambous, “Average-weight-controlled bin-oriented heuristics for the one-dimensional bin-packing problem,” European Journal of Operational Research 210, no. 2 (2011): 176-184.

[15] K. Jansen, and S. Öhring., “Approximation algorithms for time constrained scheduling,” Information and Computation 132, no. 2 (1997): 85-108.

[16] L. Epstein, and A. Levin, “On bin packing with conflicts,” In Approximation and Online Algorithms, pp. 160-173. Springer Berlin Heidelberg, 2006.

[17] B. McCloskey, and A. J. Shankar, “Approaches to bin packing with clique-graph conflicts,” Computer Science Division, University of California, 2005.

[18] B. Codenotti, G. D. Marco, M. Leoncini, M. Montangero, and M. Santini, “Approximation algorithms for a hierarchically structured bin packing problem,” Information processing letters 89, no. 5 (2004): 215-221.

[19] C. Bujtás, G. Dósa, C. Imreh, J. Nagy-GYörgy, and Z. Tuza., “The graph-bin packing problem,” International Journal of Foundations of Computer Science 22, no. 08 (2011): 1971-1993.

[20] Klein, P. Nathan, “Parallel algorithms for chordal graphs,” Brown University, Department of Computer Science, 1991.

[21] E. Falkenauer, “A hybrid grouping genetic algorithm for bin packing,” Journal of heuristics 2, no. 1 (1996): 5-30.

[22] JV. De Carvalho, “LP models for bin packing and cutting stock problems,” European Journal of Operational Research. 2002 Sep 1;141(2):253-73.

[23] P. Schwerin, and G. Wäscher, “The bin-packing problem: A problem generator and some numerical experiments with FFD packing and MTP,” International Transactions in Operational Research 4, no. 5-6 (1997): 377-389.

[24] G. Wäscher, and T. Gau, “Heuristics for the integer one-dimensional cutting stock problem: A computational study,” Operations-Research-Spektrum 18, no. 3 (1996): 131-144.

[25] G. Belov, and G. Scheithauer, “A branch-and-cut-and-price algorithm for one-dimensional stock cutting and two-dimensional two-stage cutting,” European journal of operational research 171, no. 1 (2006): 85-106.

[26] A. Scholl, R. Klein, and C. Jürgens, “Bison: A fast hybrid procedure for exactly solving the one-dimensional bin packing problem,” Computers & Operations Research 24, no. 7 (1997): 627-645.

[27] A. C. Alvim, C. C. Ribeiro, F. Glover, and D. J. Aloise, “A hybrid improvement heuristic for the one-dimensional bin packing problem,” Journal of Heuristics 10, no. 2 (2004): 205-229.

[28] S. Albers, and H. Matthias, “Semi-online scheduling revisited,” Theoretical Computer Science 443 (2012): 1-9.

[29] T. H. Cormen, C. E. Leiserson, and R. R. Rivest, Introduction to Algorithms. MIT Press (1990).

TABLE I. COMPARISON BETWEEN PROPOSED HEURISTICS A1, A2 AND FOUR EXISTING HEURISTICS WITH 1615 INSTANCES

	(a)			**(b)**			**(c)**			**(d)**			**(e1)**			**(e2)**			**(e3)**		
	Best	**Avg. Dev.**	**Max Dev.**	**Best**	**Avg. Dev.**	**Max Dev.**	**Best**	**Avg. Dev.**	**Max Dev.**	**Best**	**Avg. Dev.**	**Max Dev.**	**Best**	**Avg. Dev.**	**Max Dev.**	**Best**	**Avg. Dev.**	**Max Dev.**	**Best**	**Avg. Dev.**	**Max Dev.**
FFD	6	6.73	24	0	1.04	2	3	0.82	1	5	0.82	1	546	0.39	5	236	1.56	21	0	3.4	4
B2F	41	1.23	4	64	0.68	1	4	0.76	1	4	1.07	2	639	0.13	3	363	0.44	9	7	0.3	1
MBS	40	0.98	3	33	0.84	1	13	0.24	1	1	3.32	10	252	1.47	3	387	0.27	5	0	2.6	3
MBS´	41	1.24	7	36	0.82	1	13	0.24	1	2	1.39	3	633	0.14	3	381	0.34	6	0	3.3	4
A1	31	1.45	7	33	0.84	1	8	0.53	1	2	1.43	3	624	0.16	3	367	0.40	6	0	3.4	4
A2	23	1.28	4	75	0.63	1	8	0.53	1	5	0.82	1	631	0.15	3	381	0.34	6	4	0.6	1

An Approach of nMPRA Architecture using Hardware Implemented Support for Event Prioritization and Treating

Ionel ZAGAN[1,2], Nicoleta Cristina GAITAN[1,2] and Vasile Gheorghita GAITAN[1,2]
[1]Stefan cel Mare University of Suceava, 720229, Romania
[2]Integrated Center for Research, Development and Innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for Fabrication and Control (MANSiD), Stefan cel Mare University, Suceava, Romania

***Abstract*—One of the fundamental requirements of real time operating systems is the determinism of executing critical tasks and treating multiple periodic or aperiodic events. The present paper presents the hardware support of the nMPRA processor (Multi Pipeline Register Architecture) dedicated to treating time events, interrupt events and events associated with synchronization and inter-task communication mechanisms. Because in real time systems the treatment of events is a very important aspect, this paper describes both the mechanism implemented in hardware for prioritizing and treating multiple events, and the experimental results obtained using Virtex-7 FPGA circuit. The article's element of originality is the very short response time required in treating and prioritizing events.**

***Keywords*—nMPRA; event treating; mutex; inter-task communication; hardware scheduler**

I. Introduction

Context switching and treating periodic and aperiodic events represent key factors in implementing real time schedulers, because they enable the operating system to allocate immediately higher priority events to the processor. In full-preemptive systems, the execution of the current task can be interrupted at any time by a task with a higher priority. In some implementations used for embedded systems, context switching can be completely prohibited in order to avoid unpredictable interferences between tasks, but also for enhancing the system's predictability. For some real time systems (RTS), the preemptive scheduler can be disabled only for certain periods of time during the execution of critical sections, such as the ISR (Interrupt Service Routines).

In these days, most commercial operating systems for embedded systems do not allow a task to synchronize with more events used for resource sharing, time management or asynchronous interrupts treatment.

A parameter with a negative influence on the performance on a real-time system is the over-control due to the operating system [1]. The scheduling algorithm and task context switching operations may significantly influence the scheduling limit for those systems with a high frequency of task switching. In many practical situations, such as I/O scheduling, or communication using shared environments, an interrupt is hard, or even impossible, to accept. This is because suspending the current task would cause an increase of the cache miss effect and negatively influence the pre-fetch mechanism, by involving an unpredictable worst-case execution time (WCET).

For identifying the peripheral device that generated the interrupt, four types of techniques can be employed [2]. The existence of multiple interrupts lines between the CPU and the I/O modules is the simplest one. Therefore, even if multiple lines are used, each line is likely to have attached more than one I/O module, so one of the following three techniques could be used for each line. The software pool technique lies in the fact that when the CPU detects an interrupt, it performs a branch to the ISR that tests each I/O module in order to determine which module triggered the event. The major disadvantage of this method is that it is inefficient and time consuming. Daisy chain, also called vectored interrupt, is a more efficient technique implemented in hardware and based on the recurrent checking of interrupt signals (hardware pool). In case of treating interrupts by using this method, all I/O modules partake the same line interrupt request. Bus arbitration is another technique for treating interrupts that uses the vectored interrupts concept. In this case, a priority scheme must be used, so that the process of assigning priorities to multiple devices deals with situations in which multiple I/O modules aim simultaneously at taking control.

In order to address questions and problems related to current RTS, this paper validates the treatment of multiple events by the nMPRA processor [3], thus demonstrating the functionality and the real time performances of the integrated scheduler and the flexibility of the nMPRA processor.

This paper is structured as follows: the first section contains a brief introduction, and section two describes the nMPRA processor architecture; section III addresses the implementation in hardware of the mechanism of treating multiple events; the experimental results thus obtained will be analyzed and discussed in section IV; section V presents related work and the paper ends with the final conclusions in section VI.

II. nMPRA Architecture and Hardware Support

The nMPRA processor implemented for n threads is based on a hardware implemented scheduler as an integral part of the processor entitled Hardware Scheduler Engine (nHSE). The

Sectoral Operational Program for Increase of the Economic Competitiveness co-funded from the European Regional Development Fund.

nMPRA processor based on the five stage pipeline assembly line is designed to execute the MIPS instruction set [4], implementing new instructions for task scheduling operations. The nMPRA concept replaces the stack saving classical method with a remapping technique, which uses the replication of program counter, register file and pipeline registers for n threads, as shown in Figure 1. So, an instance of the CPU will be called semi CPU (sCPUi for task i). In order to implement the nMPRA project, the authors use nHSE with static scheduling algorithms for tasks, interrupts, and events. nHSE is directly responsible for the remapping operation of the pipeline registers set and of the registers file [5].

In order to ensure a rapid context switching, the nMPRA architecture is based on multiplying the general purpose registers. Thus, each semiprocessor implements 32 registers on 32 bits representing a bank of registers, and all banks make the register file of the nMPRA architecture. The control unit generates the control signal for the register file, according to the decoded instruction and the active sCPUi, so that, at a given time, three simultaneous operations are allowed for the same bank of registers, one for writing and two for reading. The register bank is switched through the signal generated by the nHSE module and each sCPUi has a corresponding register bank. The function contexts saving and the parameters transmission are performed in a similar way to that in the case of classic MIPS processors. The bank selection is performed in hardware, the operation being independent of instructions executed at the level of each sCPUi [6].

The ID/EX pipeline register stores the data signals obtained from decoding the instruction and extracting the operands from the register file and the control lines needed in the following stages. Therefore, in the ID stage from the assembly line, the decoding of the instruction read from memory in the IF stage and the reading of data from the register file are performed. The operands read from the register file will be either stored in the next pipeline stage, if the MIPS instruction is type R or I, or ignored as in the case of jump instructions [7]. The shift registers for memory alignment in a Word-wide memory of 32 bits is designed in the ID pipeline stage, and, in order to ensure the width of the word data, the hardware support for sign extended operation is also designed. For accessing the data memory during the reading or writing operations, the memory controller has been implemented in the MEM pipeline stage.

The control units dictate the operations of reading and writing in the data memory through the M_MemRead and M_MemWrite control signals; the control signals are transmitted and stored at every clock cycle once with the instruction context, throughout all stages of the assembly line. In designing these processor architectures, the operations of reading and writing data in memory are performed during a clock cycle, the on-chip implemented memory being dual-port, with multiple access that runs at a superior frequency to that of the processor. Because the current implementation places special emphasis on the development and validation of the nMPRA processor, ensuring the predictable execution of tasks in a hard real-time system with mixed-criticality, the memory controller and on-chip memory were designed only to meet the resource requirements for the validation of the nMPRA project.

The WB stage performs the writing of the result in the register or in the subsequent stage when the hazard detection signals the emergence of a hazard situation. Being previously memorized in the MEM/WB pipeline register, the multiplexer from the WB stage, controlled by the WB_MemtoReg control signal provided by the control unit, performs the selection of the registers resulting from the ALU unit and of the data read from memory. According to the arithmetic or logical operation, or the access to memory performed by the instruction executed, this stage will provide at output the necessary data.

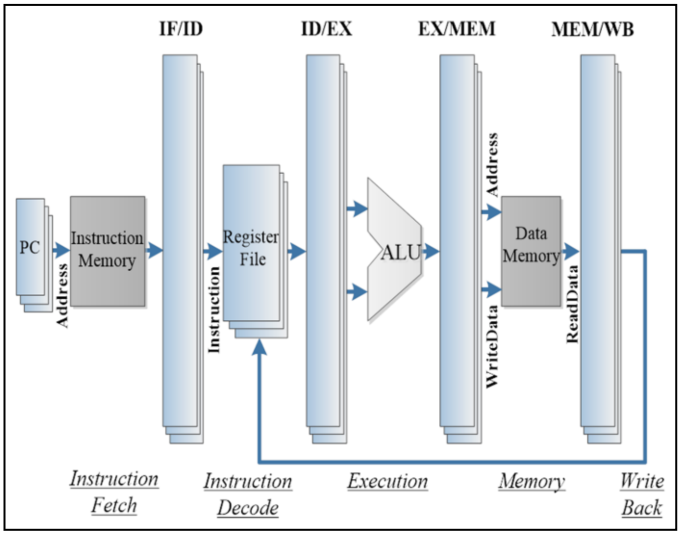


Fig. 1. Replication of resources of the nMPRA architecture. PC-program counter, IF/ID-Instruction Fetch/Instruction Decode, ID/EX-Instruction Decode/EXecute, EX/MEM-Execute/MEMory, MEM/WB-MEMory/Write Back pipeline register [8]

In the case of classical processor architectures, the saving and restoring of contexts is achieved through operations of accessing the external memory; the time needed to perform these operations depends directly on the number and the dimensions of the saved registers and the width of the data bus between the processor and the RAM memory.

The nMPRA processor uses a Harward memory architecture and the access to data and to instructions is performed in a separate address space. For the module to access the dual-port memory, the interface for both data and addresses is on 32 bits, using the big endian or little endian format, depending on the value of the Big_Endian parameter, set in the MIPS_Parameters.v file. The nMPRA supports memory accesses of type word, halfword and byte. The data memory bus is synchronous, used to access the RAM on-chip memory. It uses a minimum number of control signals and a simple protocol, in order to ensure that the data and instruction memory is accessed in writing, in a single clock cycle; the access to memory is performed on the positive edge of the clock signal.

III. Prioritizing and Treating Events by the nHSE Scheduler

The nHSE is a finite state machine which has inputs for events, such as interrupts, deadline, watchdog timers, timers, mutexes, messages, and self-support execution. This implementation allows a very fast context switching that is possible due to the remapping of the active running task context with the scheduled task; the jitter is minimized in order to provide an accurate predictability behavior.

The nHSE architecture implements a hardware block with the role of arbitrating and announcing the sCPUi that has attached the event in question either directly, or through the active scheduler (static or dynamic). This block validates the command signals for each sCPUi. The events representing input signals for the nHSE module are the following: interrupts, the timer generated event, the event generated by exceeding deadline 1, the event generated by exceeding deadline 2, the event generated by the watchdog timer, the events generated by mutexes and the synchronization events.

The dynamic scheduler represents the support for the dynamic scheduling that enables the priority switching of a task, including that of a sCPUi. This scheduler is deactivated on reset only by the sCPU0. The nHSE module contains one register with the identifier corresponding to each sCPUi, one register with the priority set for the sCPUi used only by the dynamic scheduler, and a global register containing the identifier for the active sCPU that can be inhibited during the execution of atomic instructions. The sCPU0 will aways have the highest priority that is priority 0.

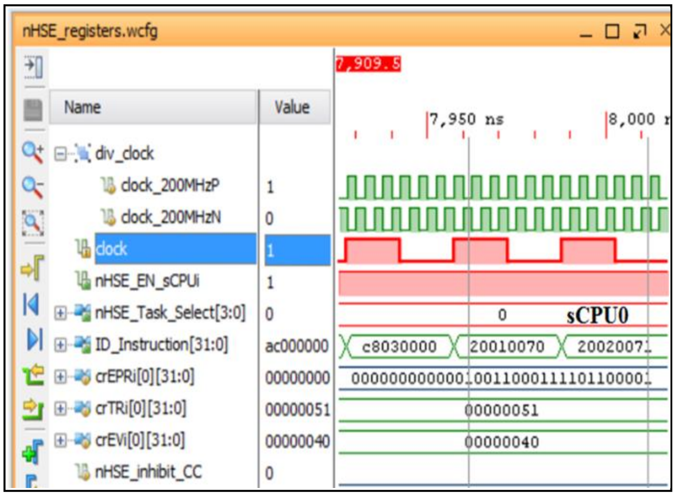
The crEPRi[i] register presented in Table 1, represents the priorities attached to each event that can be validated or not at the level of each sCPUi. Thus, each sCPUi can have different priorities for time events, interrupts, mutexes and synchronization events through messages.

The Pri_TEvi, Pri_WDEvi, Pri_D1Evi, Pri_D2Evi, Pri_IntEvi, Pri_MutexEvi and Pri_SynEvi bits groups represent the priorities attached to the categories of events validated or inhibited in the crTRi control register. Thus, when an event occurs, the corresponding bit will be set from the crEVi register.

The fact that interrupts have fixed priorities should be emphasized; the grINT_IDi[0] interrupt has the highest priority and the grINT_Idi[p-1]interrupt has the lowest; p is the number of interrupts from nMPRA. Although the priority of interrupts is fixed [9], they can be attached to any sCPUi and, at the level of each sCPUi, they can have different priorities given by the priorities set in the crEPRi register. The selection of the interrupt with the highest priority is performed through a hardware module that implements the priority encoder for interrupts.

IV. Experimental Results

The project has been implemented using the VC707 Evaluation Kit produced by Xilinx and Vivado 2015.4 design environment and the source code has been written in Verilog HDL. The implementation is based on the project described in [10], a 32-bit MIPS processor which aims for conformance with the MIPS32 Release 1 ISA. Figure 2 shows the clock_200MHzP and clock_200MHzN clock signals which represent the 200MHz differential signal available at the output of the SIT9102 oscillator and the clock signal of the nMPRA processor (clock) generated through the PLL block obtained with IP Clockind Wizard 5.2 (Rev. 1).


nHSE_registers.wcfg

Name	Value
div_clock	
clock_200MHzP	1
clock_200MHzN	0
clock	1
nHSE_EN_sCPUi	1
nHSE_Task_Select[3:0]	0
ID_Instruction[31:0]	ac000000
crEPRi[0][31:0]	00000000
crTRi[0][31:0]	00000051
crEVi[0][31:0]	00000040
nHSE_inhibit_CC	0

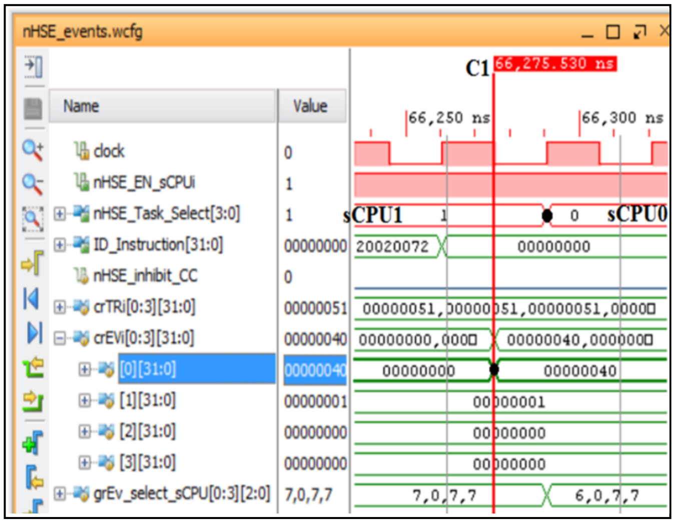
Fig. 2. The registers of the nHSE hardware integrated scheduler

The nMPRA processor architecture, using Virtex7 development kit, is defined and validated in the present paper, without describing the entire SoC project. Particular attention was paid to the nHSE real time scheduler, to improving the execution predictability by partially or completely eliminating hazards from the pipeline and minimizing the jitter for task context switching [11][12][13][14]. Compared to the theoretical version, in the version used to validate the processor, two clock signals were used. One clock cycle was used both for the pipeline registers, the register file, the internal logic of the scheduler and for treating external asynchronous interrupts [15]. The second clock cycle was used for the instruction and data memory.

The waveforms corresponding to the nHSE_EN_sCPUi, nHSE_Task_Select[3:0], ID_Instruction[31:0], crEPRi[0][31:0], crTRi[0][31:0], crEVi[0][31:0] and nHSE_inhibit_CC signals are also represented. The nHSE module generates the activation signals for all sCPUi semiprocessors through the nHSE_Task_Select[3:0] selector and the nHSE_EN_sCPUi validation signal; it can be inhibited under certain conditions, by the logic of the n events.

TABLE I. Assigning Priorities to Multiple Events using the crEPRi Control Register

31..21	20..18	17..15	14..12	11..9	8..6	5..3	2..0
-	Pri_MutexEvi	Pri_MutexEvi	Pri_IntEvi	Pri_D2Evi	Pri_D1Evi	Pri_WDEvi	Pri_TEvi



Name	Value
clock	0
nHSE_EN_sCPUi	1
nHSE_Task_Select[3:0]	1
ID_Instruction[31:0]	00000000
nHSE_inhibit_CC	0
crTRi[0:3][31:0]	00000051
crEVi[0:3][31:0]	00000040
[0][31:0]	00000040
[1][31:0]	00000001
[2][31:0]	00000000
[3][31:0]	00000000
grEv_select_sCPU[0:3][2:0]	7,0,7,7

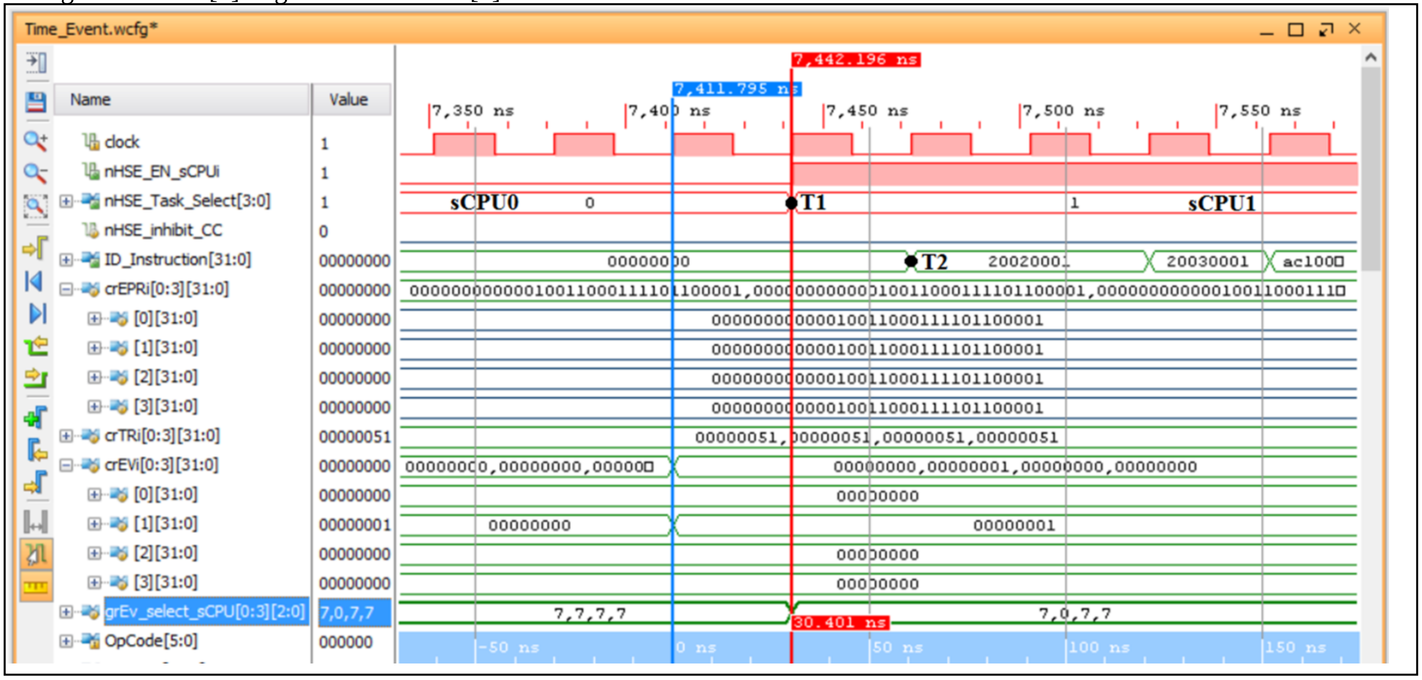
Fig. 3. The nHSE scheduler dictates a context switch for sCPU0 to treat the the synchronization event through messages

The grEv_select_sCPU[0:3][2:0] registers store the events treated by each sCPUi at any time moment. As can it be seen in Figure 3, the moment marked by the cursor C1 indicates the occurrence of a synchronization event through messages; the event is stored by the crEVi[0] = 0x00000040 register. Even if the sCPU1 semiprocessor treats a time event (grEv_select_sCPU[1] = 0), the scheduler performs the context switching operation because sCPU0 has the highest priority and the nHSE_inhibit_CC signal does not prevent the switching of the semiprocessors. The value 6 stored in the grEv_select_sCPU[0] register indicates the fact that sCPU0 treats a synchronization event through messages, validated through the crTRi[0] register. The crTRi[0] = 0x00000051 register validates time events, interrupt generated events, and events generated by the mechanism of communicating through messages.

As we can see in Figure 3, the contexts switch operation is guaranteed in one clock cycle. In a four sCPUi version as the one used for obtaining the waveforms in the present article, we can observe the ID_Instruction[31:0] pipeline register containing, at a certain moment, the code for the instructions extracted for each sCPUi.

Figure 4 represents the situation when there is a time event; time moment T1 represents the moment in which context switching is performed; the data stored in the pipeline registers are saved during the transition from sCPU0 to sCPU1. At time moment T2, the first instruction corresponding to sCPU1 (ID_Instruction[31:0] = 0x20020001) is extracted. This switch takes place under the strict command of the nHSE static scheduler, through the nHSE_Task_Select[3:0] and nHSE_EN_sCPUi nHSE signals, the time needed for switching contexts is no more than one clock cycle.

The crEPRi[0:3][31:0] register stores the priorities of the 7 events validated through the crTRi[0:3][31:0] registers. Thus, at the level of the semiprocessor sCPU0 corresponding to the validated events, the following priorities are already set: Pri_TEvi=3’b001, Pri_IntEvi=3’b000 and Pri_SynEvi=3’b010. A smaller value represents a higher priority, Pri_IntEvi=3’b000 being the event with the highest priority. The priority level of each category of events can be changed dynamically through the instructions dedicated to the nHSE scheduler, in relation to the requirements of the real time system.



Name	Value
clock	1
nHSE_EN_sCPUi	1
nHSE_Task_Select[3:0]	1
nHSE_inhibit_CC	0
ID_Instruction[31:0]	00000000
crEPRi[0:3][31:0]	00000000
[0][31:0]	00000000
[1][31:0]	00000000
[2][31:0]	00000000
[3][31:0]	00000000
crTRi[0:3][31:0]	00000051
crEVi[0:3][31:0]	00000000
[0][31:0]	00000000
[1][31:0]	00000001
[2][31:0]	00000000
[3][31:0]	00000000
grEv_select_sCPU[0:3][2:0]	7,0,7,7
OpCode[5:0]	000000

Fig. 4. The treating of an event using the nHSE architecture; clock - nMPRA clock; nHSE_EN_sCPUi - nHSE enable signal; nHSE_Task _Select[3:0] - nHSE task selector; nHSE_inhibit_CC – context switch inhibit signal; ID_Instruction[31:0] - wire type instruction; crTRi – enable event register; crEVi – events register; grEv_select_sCPU - current event identifier

The grINT_PR global register implemented in the nHSE scheduler stores the number of the interrupt with the highest priority, the selection being performed in hardware. The nMPRA architecture guarantees the execution of the new scheduled task starting with the next clock cycle, as we can see in Figure 4, at the moment T1.

We remind that all sCPUi share the same functional units, such as ALU, the control unit, the condition unit, the unit for hazard detection, and the redirection of data unit, so that the data path must guarantee the hardware isolation and the consistency of sCPUi contexts [16]. At a 33MHz frequency, the scheduler answer to an time related event may be around 30.401ns (one clock cycle). It can be said that the experimental results demonstrate the practical implementation of the theoretical aspects, therefore obtaining very low times for events handling and context switching operations.

The aim of this test is to verify and validate the custom interrupt management scheduling policy implemented in nHSE, and emphasize the added performance brought by the nMPRA processor in comparison to the theoretical elements presented in section III. Because the datapath is shared by all sCPUi implemented in the nMPRA processor, their contexts must be preserved and made available at any time moment to the real-time nHSE scheduler.

The goal of this implementation is not to describe a complete solution of the data path, but to validate the practical implementation of the nMPRA architecture and of the nHSE scheduler, using a flexible and competitive FPGA development platform. To design the nHSE module and to obtain better performances brought by the nMPRA, an analysis of events handled through software as well as hardware was necessary in the case of treating interrupts in a classical computing system.

V. Related Work

This chapter presents a brief description of two predictable processor architectures which can be compared with the results presented in this paper using the nMPRA processor.

The Merasa concept [17] was developed to obtain a processor architecture which can be successfully used in hard real-time embedded systems. Concerning the architecture, each core can have only one hard real-time (HRT) execution thread and an arbitrary number of non-HRT (NHRT) execution threads. Each core is made up of two scratchpad memories; one of the memories is dedicated for data and the other for instructions (D-ISP and DSP); the data integrity is ensured by individual allocation of a subnet of banks cache for each task. The priorities for execution threads are fixed and the scheduling policy chosen is round robin. Taking into consideration that the embedded systems have limited resources available, the Merasa architecture must offer an optimal cost for the implementation of an average number of HRT and NHRT execution threads, including their synchronization and communication mechanisms. If the HRT thread is suspended, pending an external interrupt, time event, or sharing a resource with another HRT or NHRT task, its dedicated assembly line will remain unused and it will negatively influence the performance of the entire system.

In [18], Andalam proposes a new predictable architecture called ARPRET. The ARPRET architecture is implemented and synthesized on the Xilinx ML-403 FPGA device, obtaining predictability by projecting a particular soft-core coupled with a hardware accelerator, called the Predictable Functional Unit. Thus, time behavior for models and programs becomes most important because, in order to guarantee that a hard real-time system behaves according to the model, their characteristics must be preserved during compilation.

VI. Conclusion and Future Work

The nMPRA architecture is versatile and very flexible because of the following reasons: the prioritization of multiple events attached to a sCPUi, the wait instruction which enables the implementation in hardware of a logical OR between these events, and the implementation in hardware of synchronization and inter-task communication mechanisms. However, the priorities of tasks, interrupts, and synchronization mechanisms can be ordered in any way, in order to meet the requirements of the real time application, whose central element can successfully be the nMPRA processor.

Each task is executed based on its own context, without depending on other system tasks or scheduler actions. For inter-task communication, nMPRA can ensure the implementation of queues, enabling data to be safely transferred between tasks. The hardware implementation of queues is flexible and can be used to obtain a number of objectives, including simple data transfers and synchronization through mutexes. By sending and receiving data using queues, the events specific to the queues implemented in classical CPU architectures can be used.

The performances of the nMPRA architecture can be improved by designing a cache memory for optional data and a memory protection module for unauthorized access, thus satisfying the constraints of the hardware isolation of tasks. The following papers should consider both the improvement of the memory architecture and the comparison with other similar implementations. The negative collateral effect generated by obtaining these outstanding performances, essential for mixed criticality real-time systems, is the memory consumption for multiplying in hardware the multiplexed resources, such as PC, register file and pipeline registers.

Acknowledgment

This work was partially supported from the project "Integrated Center for research, development and innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for fabrication and control", Contract No. 671/09.04.2015, Sectoral Operational Program for Increase of the Economic Competitiveness co-funded from the European Regional Development Fund.

This paper has been prepared with the financial support of the project BG58/2016: "Development and integration of a mobile tele-electrocardiograph within the monitoring and diagnosis system of patients GreenCARDIO", acronym: m-GreenCARDIO, PNCDI III, P2 - increasing competitiveness through RDI Romanian economy, Transfer of knowledge to the economic agent. Contract no. PN-III-P2-2.1-BG-2016-0463.

References

[1] G. C. Buttazzo, ”Hard Real-Time Computing Systems - Predictable Scheduling Algorithms and Applications,” Third edition, pp. 13–30, Springer, 2011. ISBN: 978-1-4614-0675-4

[2] W. Stallings, “Computer Organization and Architecture,” 10th Edition, pp. 263–272, 2015. ISBN: 978-0134101613

[3] V. G. Gaitan, N. C. Gaitan, and I. Ungurean, “CPU Architecture Based on a Hardware Scheduler and Independent Pipeline Registers,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, no. 9, pp. 1661–1674, Sept. 2015. doi:10.1109/TVLSI.2014.2346542

[4] D. A. Patterson and J. L. Hennessy, “Computer Organization and Design, Revised Fourth Edition: The Hardware-Software Interface,” Fourth Edition, pp. 330–379, 2011. ISBN: 978-0-12-374750-1

[5] I. Zagan, “Improving the performance of CPU architectures by reducing the Operating System overhead,” in 3rd IEEE Workshop on Advances in Information, Electronic and Electrical Engineering AIEEE’2015, Riga, Latvia, pp. 1–6, 13–14 Nov. 2015. doi: 10.1109/AIEEE.2015.7367279

[6] E. Dodiu and V. G. Gaitan, “Custom designed CPU architecture based on a hardware scheduler and independent pipeline registers – concept and theory of operation,“ in IEEE EIT International Conference on Electro-Information Technology, Indianapolis, USA, pp. 1–5, May 2012. doi:10.1109/EIT.2012.6220705

[7] “MIPS® Architecture For Programmers Volume I-A: Introduction to the MIPS32® Architecture,” Revision 3.02, Mar. 2011, Available: https://courses.engr.illinois.edu/cs426/Resources/MIPS32INT-AFP-03.02.pdf. (Accessed: May 2016).

[8] I. Zagan and V. G. Gaitan, “Predictable CPU Architecture Designed for Small Real-Time Applications – Implementation Results,” in: 3rd International Conference on Advances in Computing, Electronics and Communication (ACEC), 10 - 11 October 2015 / Zurich, Switzerland, pp. 143-150, ISBN: 978-1-63248-064-4. doi:10.15224/ 978-1-63248-064-4-29.

[9] S. Kelinman and J. Eykholt, “Interrupts as threads,” ACM SIGOPS Operating Syst. Rev., vol. 29, no. 2, pp. 21–26, Apr. 1995. doi:10.1145/202213.202217

[10] I. Zagan and V. G. Gaitan, “Improving the Performances of the nMPRA Processor using a Custom Interrupt Management Scheduling Policy,” in Advances in Electrical and Computer Engineering (AECE), ISSN 1582-7445, Volume 16, Issue 4, pp. 45-50, 2016. doi:10.4316/AECE.2016.04007

[11] E. E Moisuc, A. B. Larionescu, and V. G. Gaitan, “Hardware Event Treating in nMPRA,” in 12rt International Conference on Development and Application Systems – DAS, Suceava, Romania, pp. 66-69, 15–17 May, 2014. doi:10.1109/DAAS.2014.6842429

[12] E. Dodiu, V. G.Gaitan, and A. Graur, “Custom designed CPU architecture based on a hardware scheduler and independent pipeline registers – architecture description”, in IEEE 35’th Jubilee International Convention on Information and Communication Technology, Electronics and Microelectronics, Croatia, pp. 859-864, 24 May 2012. INSPEC Accession Number: 12865464

[13] L. Andries and V. G. Gaitan, “Dual Priority Scheduling algorithm used in the nMPRA Microcontrollers,” in 18th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, Romania, pp. 43-47, 2014. doi:10.1109/ICSTCC.2014.6982388

[14] L. Andries, V. G. Gaitan and E. E. Moisuc, “Programming paradigm of a microcontroller with hardware scheduler engine and independent pipeline registers - a software approach,” in 19th International Conference on System Theory, Control and Computing (ICSTCC), Cheile Gradistei, Romania, pp. 705-710, 2015. doi: 10.1109/ICSTCC.2015.7321376

[15] I. Zagan and V. G. Gaitan, “Schedulability Analysis of nMPRA Processor based on Multithreaded Execution,” in 13rt International Conference on Development and Application Systems – DAS, Suceava, Romania, pp. 130-134, May 19–21, 2016. doi:10.1109/DAAS.2016.7492561

[16] N. C. Gaitan, I. Zagan, and V. G. Gaitan, “Predictable CPU Architecture Designed for Small Real-Time Application - Concept and Theory of Operation,” International Journal of Advanced Computer Science and Applications, vol. 6, no. 4, 2015. doi: 10.14569/IJACSA.2015.060406

[17] T. Ungerer et al., Merasa: Multicore execution of hard real-time applications supporting analyzability, IEEE, Micro, vol. 30, no. 5, pp. 66–75, 2010. doi: 10.1109/MM.2010.78

[18] S. Andalam, Predictable platforms for safety-critical embedded systems, Thesis, The University of Auckland, 2013.

Concepts and Tools for Protecting Sensitive Data in the IT Industry: A Review of Trends, Challenges and Mechanisms for Data-Protection

Omar Tayan
Dept. of Computer Engineering, College of Computer Science and Engineering (CCSE),
IT Research Center for the Holy Quran and Its Sciences (NOOR),
Taibah University, Saudi Arabia

***Abstract*—Advancements in storage, dissemination and access of multimedia data content on the Internet continues to grow at exponential rates, while individuals, organizations and governments spend huge efforts to exert their fingerprint in this information age through the use of online multimedia resources to propagate thoughts, services, policies, ecommerce and other types of information. Furthermore, information at different levels may be classified into confidential, sensitive and critical data types. Such data has been subject to numerous tools and techniques for providing automated information processing, information management and storage mechanisms. Consequently, numerous security tools and techniques have also emerged for the protection of data at the various organizational levels and according to different requirements. This paper discusses three important types of information security aspects that includes; data-storage, in-transit data and data access-prevention for unauthorized users. In particular, the paper reviews and presents the latest trends and most common challenges in information security with regards to data-breaches and vulnerabilities found in industry today using simple brief summaries for the benefit of IT practitioners and academics. Thereafter, state-of-the-art techniques used to secure information content commonly required in applications-software, in-house operations software or websites are given. Mechanisms for enhancing data-protection under the given set of challenges and vulnerabilities are also discussed. Finally, the importance of using information security policies and standards for protecting organizational data content is discussed along with foreseeable open issues for future work.**

Keywords—sensitive-data; data-breaches; data-protection; trend analysis; classification

I. Introduction

The Digital era has witnessed an ever increasing dependence on the Internet and world-wide web (WWW) in our lives and daily activities. Moreover, the continuing growth of such information and communication technologies has played a crucial role in establishing the Internet and WWW as the dominant IT platform for digital content distribution, communication, and other general information sharing activities. Hence, millions of worldwide users have benefited from the advantages of fast and simple mechanisms for digital information exchange. On the contrary, such benefits are also vulnerable to the problems and threats associated with securing the digital content. The literature of digital multimedia content has identified a number of security issues to be addressed that includes: digital copyright protection, counterfeit prevention and data-authentication. Such requirements are more predominant in the case of specialized and sensitive data. Generally, all digital multimedia content on the Internet can be classified into text, images, audio and video content with the challenge being to provide secure, robust and reliable storage and dissemination for each media type. On the other hand, many electronic-transactions impose an additional requirement to ensure data-confidentiality, particularly for the case of sensitive customer and client information. This paper explains the important and timely role of information assurance and related security techniques concerned with the storage, propagation, reproduction and communication of sensitive online data-content.

Background Concepts in Information Security

Some of the key objectives of digital multimedia security can be classified into; requirements for assuring authenticity and integrity of content, usage-control, binding of identification data with the cover-content, and ensuring secrecy and non-repudiation in the transmitted content [1 - 3]. The state-of-the-art techniques in information security can be used to achieve the necessary security requirements according to the target application and content-type in most cases. The protection of sensitive digital multimedia content can be achieved using authenticity and integrity based techniques to ensure that 100% accurate content is transmitted and stored, whereas secrecy of the data can be achieved using cryptographic approaches prior to transmission. *Integrity* is concerned with ensuring that the transmitted data is not altered or tampered with, and is exactly similar to the version sent. Integrity can be achieved using numerous techniques such as; encryption, hashing, watermarking etc. *Authentication*, on the other hand, is associated with establishing trust between communicating parties, such as assurance by verifying that the data-content had originated from a trusted source/publisher. Authentication can be achieved using digital signatures/certificates and digital-watermarking. In contrast, *confidentiality* and *non-repudiation* requirements are typically used with e-transactions that are concerned with data-secrecy during the communication and are achieved using encryption schemes.

II. Vulnerability Trends in the IT Sector and a Classification of Challenges for Data-Protection

Organizational employees with access to networked-devices have a key role in protecting the organization's information assets since those devices can provide a gateway to information stored elsewhere on the same network and can be exploited as vulnerable access-points for internal or external intruders. In fact, an organization faces a number of risks due to many types of possible information security vulnerabilities, which typically include:

- Fraudulent websites that can imitate other sites
- Data-theft
- Fake purchases
- Intruder attacks
- Damage to an organization's reputation

Moreover, this information-era has witnessed many ways in which data and security-breaches have penetrated our normal business operations and daily-life activities. Such security-breaches can now be found in most/all IT systems covering new and known application-domains and functions, including: e-Banking and e-Commerce applications [4], e-Healthcare systems [5], wireless and mobile devices [6,7], cloud-assisted applications, wireless sensor-networks (WSN) and Internet-of-Things (IoT) [8] and Big-Data processing activities [9]. Figure 1 illustrates those recent domains with emerging penetrations due to security-breaches [4-9].

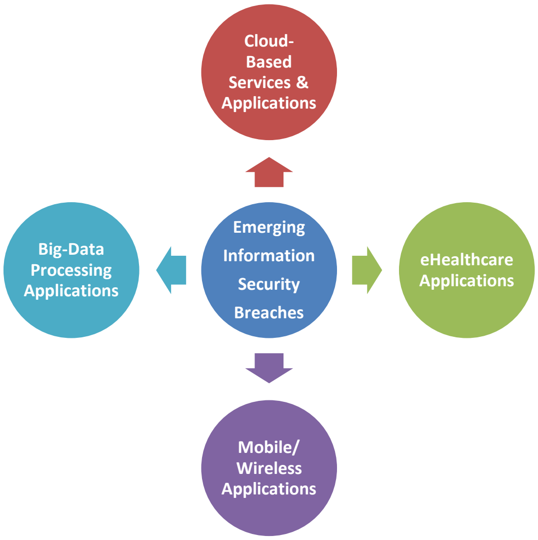

Cloud-Based Services & Applications
Big-Data Processing Applications
Emerging Information Security Breaches
eHealthcare Applications
Mobile/ Wireless Applications

Fig. 1. Emerging Information Security Breaches by Domain

Table 1 classifies the actual extent of damage faced by organizations and individuals due to information security vulnerabilities and cyber-attacks and includes threats/vulnerability statistics industrial-sector:

TABLE I. Classification of Cyber-Attacks and Victim Counts in 2015 [10]

Category	Attack/Vulnerability Classification	Quantitative Analysis
Personal Digital Identities	Personal Identities Lost or Stolen	Over Half a Billion
Data-Breaches	Total Breaches	318
	Average Num of Identities Exposed/Breach	1.3 Million
	Total Identities Exposed	429 Million
	Top Causes of Data Breaches by Identities	Attackers, Accidental-loss, Theft
	Top Ranked Industry Sectors Targeted for Data-Breaches	Services (Inc. Healthcare), Finance, Public Admin., Wholesale Trade, Retail Trade
Email Threats, Malware and Bots	Overall Email/Spam Ratio	53% of emails
	Email Phishing Ratio	1 in 1846 emails
	Email Malware Ratio	1 in 220 emails
	Number of Bots	1.1 Million
	Number of New Malware Variants	431 Million
	Top-Ranked Industries Targeted by Spam Emails	Mining, Manufacturing, Construction, Services, Agriculture
Mobile Devices	New Vulnerabilities Detected	528
	New Android Malware Variants	3944
	Ratio of Apps Analyzed and Classified as Malware	Over 30%
Vulnerabilities	New Vulnerabilities	5585
	Zero-Day Vulnerabilities	54
	Most Frequently Targeted App for Zero-Day Exploit Vulnerabilities	Adobe Flash Player
Web-Attacks	Scanned Websites with Vulnerabilities	78% (15% of which were critical)
	Websites Detected with Malware	1 in 3172
	Top-Ranked Most Frequently Exploited Sites	Technological sites, Business sites, Searching sites.

Other interesting facts related to attack-frequencies, types and detection-rates with a classification of countries and regions are now summarized in Table 2.

TABLE II. STATISTICAL SUMMARY OF VARIOUS VULNERABILITIES, THREATS AND ATTACKS

Category	Reference / Citation	Summary of Statistical Trends
Infected Smartphone Apps	Mobile Threats Report [11]	150M apps scanned, with 9M malwares, 9M suspicious apps and 3M affected devices detected
Top 10 Countries Ranked by Infections for Smartphone Devices in 4th-Quarter, 2015		India (86k infections), USA (82k infections), Brazil (68K infections)
Largest Rate of Increase in Mobile Malware Threats		From Third to Fourth Quarter, 2015 (72% increase)
Total Worldwide Malwares Detected	McAfee Labs Threats Report [12]	Total of nearly 500M, with the largest growth in the fourth-quarter in 2015 due to newly emerging mobile threats (representing 12M in the fourth-quarter of 2015). Highest regional malware infection rates were reported Africa, followed by Asia and South America.
Total Worldwide Web-Threats Detected		New suspect and phishing URLs reaching 17M and 1.4M, respectively by the end of 2015, with global spam emails reaching over 4.5 trillion messages.
Top Network Attacks Detected		Browsers (36%), Brute-Force attacks (19%), DoS attacks (16%), SSL attacks (11%).

In [13-18], cyber-attacks/e-crimes were classified into three categories. The first category relates to external-attackers, which involves attacks using viruses, worms, Trojan-horses, and Denial-of-Service (DoS). Next, internal-crimes were classified as those that include unauthorized access, theft of IP-rights/theft of knowledge by employees and breach of conduct by business partners. Finally, the social-engineering category of attacks had included; phishing and spoofing. Moreover, the work in [14] provides a report on the common technical vulnerabilities in web-applications and websites, which can be summarized into the following aspects:

- Cross-site Scripting (XSS)
- SQL Injection
- Malicious File-Inclusion
- Insecure Direct Object Reference
- Information Leakage/Improper Error Handling
- Insecure Cryptographic Storage
- Insecure Communications
- Broken Authentication/Session Management
- Failure to Restrict URL Access

Numerous examples exist relating to the extent of such security breaches within the various IT-based industrial-sectors, and particularly in the case of many highly-reputable and financially-strong organizations as shown in Table 3.

TABLE III. WORLDWIDE IMPACT OF SECURITY BREACHES ON VARIOUS IT-BASED INDUSTRIES [15]

Organization/ Sector	Victim of Security Breach
Academic	Univ of Utah (2007), Univ. of Miami (2007), Stanford Uni. (2008), Univ of Calif/Berkeley (2008), Ohio State Univ. (2009), Yale Univ. (2009), Kirkwood Community College (2013), Indiana Univ. (2014).
Energy	GS Caltex (2007), New York State Electric & Gas (2011), Central Hudson Gas & Electric (2013).
Financial	Citigroup (2005), Cardsystems Solutions Inc. (2006), Ameritrade Inc. (2006), Countrywide Financial Corp (2006, 2010), Compass Bank (2007), RBS Worldpay (2008), US Federal Reserve Bank Cleveland (2009), Heartland (2009), JP Morgan Chase (2009, 2015), Citigroup (2010, 2014), Court Ventures (2011).
Government	US Dept. Of Vet. Affairs (2006), UK Revenue & Customs (2006), UK MoD (2007), UK Home Office (2007), Chile MoE (2008), US Law Enforcement (2009), Classified War Docs (2009), State of Texas (2010), Greek Government (2013), South Africa Police (2013), Florida Courts (2014).
Health/ Healthcare	Health Net (2008), Virginia Dept. of Health (2008), Affinity Health Plan Inc. (2009), NY City Health and Hosptials Corp. (2009), NHS (2010), TriCare (2010), Medicaid (2013), Advocate Medical Group (2013), Anthem (2015), Premera (2015).
Military	US Dept. of Vet Affairs (2006), US National Guard (2007), US Dept. of Defense (2008), US Military (2008, 2009), Tricare (2010), US Army (2011), Militarysingles.com (2012).
Tech./Telecoms	T-Mobile (2006), KDDI (2006), HP (2006), AT&T (2008, 2009), KT Corp. (2011), Apple (2012, 2013), Ubuntu (2012), Yahoo Voices (2012), Adobe (2013), Vodafone (2013), Yahoo Japan (2014), Terracom & YourTel (2014).
Web	AOL (2004, 2005, 2014), Monster.com (2006), RockYou (2009), China SW Developer Net (2010), Steam, Tianya, Gamigo (2010), Dropbox (2011), Zappos (2012), LinkedIn (2012), Twitter (2012), Facebook, Drupal, Scribed (2013), LivingSocial (2013), Ebay (2014), Mozilla (2015).

Notably, a number of data security breaches can also be recalled that relate to some recent and famous incidents with impact on most online users today. Some of those recent events include:

- **LinkedIn Accounts –** 6.5 million accounts were hacked on 5th June'12 and passwords publicly posted on 6th June' 2012.
- **ARAMCO attack –** 15th August 2012 – virus Shamoon attacks 30,000 PCs at company, taking Aramco two-weeks to recover.
- **Facebook –** most popular social networking site had around 600,000 "compromised" accounts/day.

Table 4 classifies the top fifteen countries involved in the generation of those attacks that had resulted with consequent data security breaches according to another study [13].

TABLE IV. TOP FIFTEEN COUNTRIES FROM WHICH DATA-BREACHES WERE GENERATED DURING THE OBSERVED-PERIOD

Source of Attack	Number of Attacks
Russia	2,402,722
Taiwan	907,102
Germany	780,425
Ukraine	566,531
Hungary	367,966
USA	355,341
Romania	350,948
Brazil	337,977
Italy	288,607
Australia	255,777
Argentina	185,720
China	168,146
Poland	162,235
Israel	143,943
Japan	133,908

III. CLASSIFICATIONS OF TECHNICAL AND ORGANIZATIONAL-LEVEL TECHNIQUES FOR PREVENTING DATA-BREACHES AND ENHANCING DATA-PROTECTION

The discussion presented in this section comprises of technical approaches, organizational approaches and strategies for managing information-security requirements, as follows:

A. *Technical Approaches*

Some of the main technical requirements concerned with the protection of sensitive content are summarized in Table 5.

TABLE V. SUMMARY OF RECURRING REQUIREMENTS AND COMMENTS FOR PROTECTING SENSITIVE DATA-CONTENT

No.	Requirements	Comments
1	Digital Information Exchange	Benefits to millions of users. Associated with problems/threats.
2	Digital Content Protection	Counterfeiting, proof-of-authentication, content-originality challenges.
3	Security Requirement/Sensitivity of Digital Multimedia	Integrity & Authentication needed. Secure from tampering.
4	Digital Watermarking	Effective security for sensitive data.

Table 6 highlights common state-of-the-art security-techniques that have emerged as a consequence, together with their goals the corresponding application-domains.

TABLE VI. CLASSIFICATION OF COMMON TECHNIQUES, THEIR GOALS AND APPLICATIONS IN INFORMATION SECURITY

Technique	Goal / Objective	Application
Encryption Systems	Confidentiality and Integrity	Symmetric-Key systems Asymmetric/Public-key systems
Watermarking, Digital Certificates	Authentication and Integrity	Adds signature of source in data Used for tracing and copyright protection
Steganography	Authentication and Integrity	Purely data-hiding purposes High-capacity embeddings
Fingerprinting, Message Digests, Hash Functions	Authentication and Integrity	Used in secure hash-algorithms, one-way hashing.
Protocols	Confidentiality and Integrity	Provides a known communication mechanism between 2+ parties
Hybrid Systems	Confidentiality and Integrity	Combines between symmetric and asymmetric key systems Session-key can be applied.

B. *3-Tier Organizational Approach*

A summary of the procedures and guidelines that forms part of an organizational action-plan for protecting digital information can be further classified into three levels (management level, implementation level and systems level) as follows:

Management Level Protection (General advice):

- Assign a Chief Security Officer (CSO).
- Develop an organizational security-policy
- Seek third-party accreditation that ensures high-security standards are achieved, e.g. ISO 27001, ISO 9001for improving quality-standards and overall reputation.
- Perform regular risk assessments and revise management solutions currently in-place.

Implementation Level Protection (summarized from [16]):

- Educate employees of the organization’s security policies.
- Raise awareness of the network-administrator/IT helpdesk role and contact details.
- Be mindful of how to share sensitive data across the network.
- Do not open unexpected email attachments or downloads.
- Perform regular backups, password-updates, encryption, biometric control.

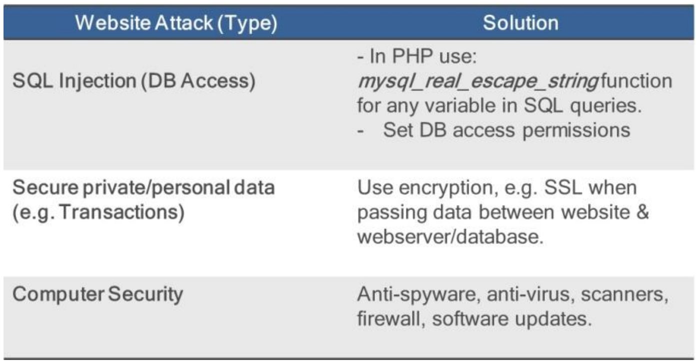
- Caution should be taken not to email content that you would not want to be distributed to unauthorized parties.
- Ensure data-sharing features on the PC are off or set to allow access to authorized persons only.
- Keep the system and security updates active and patched on PCs.
- Do not store sensitive data in an unsecure location online.
- Remote access to an organization's PCs should be done via secure methods (e.g. SSH/VPN).

Systems Level Protection (summarized from [17]):

- Select a secure e-commerce hosting platform.
- Use a secure connection for online transactions that is PCI compliant (e.g. SSL certificates).
- Do not store sensitive customer details (e.g. card numbers).
- Use address and card verification systems.
- Request customers to use strong passwords.
- Setup alert systems for suspicious activity (e.g. same IP/person may be using many card numbers).
- Use Layered Security (e.g. Perimeter, Network, Host, Application and Data layers) such as firewalls, contact-forms, and login boxes.
- Provide Security training to employees.
- Use tracking-numbers for all e-transactions or orders.
- Monitor your site regularly (e.g. use RT-analytics tools to view interaction) and ensure that the hosting platform continuously monitors their own servers (e.g. against malware, viruses, updates needed).
- Perform regular PCI scans (e.g. using Trustwave, PrestaShop).
- Patch/Update systems and third-party code (including perl, java, php, joomla, wordpress).
- Use DDoS protection service and mitigation service (e.g. Cloud DDoS protection and DNS service).
- Consider a fraud-management service from a card-company.
- Ensure the platform host regularly backs up the site and has a disaster-recovery plan.
- Encrypt stored, transmitted and processed data.

Table 7 identifies a number of quick-tip solutions for several very common web-based attacks at the system-level and implementation-level.

TABLE VII. QUICK TIPS PROVIDING SOLUTIONS TO MOST COMMON ATTACKS [18]



Website Attack (Type)	Solution
SQL Injection (DB Access)	- In PHP use: *mysql_real_escape_string* function for any variable in SQL queries. - Set DB access permissions
Secure private/personal data (e.g. Transactions)	Use encryption, e.g. SSL when passing data between website & webserver/database.
Computer Security	Anti-spyware, anti-virus, scanners, firewall, software updates.

C. Strategies for Managing Information Security

When an organization evaluates the need and extent for information security techniques against the deployment costs, a number of considerations must be made as part of a complete strategy that includes [19]:

- A chief-security officer (CSO) must balance the trade-off between risks and costs for securing the organization's assets.
- The security-management approach should consider:
  - Determining the information assets and their value
  - Determining the maximum time which the organization can function without a given asset.
  - Implementing security-procedures to protect each asset.
  - **Loss Calculations** should be used to justify costs for purchasing security techniques: *Annual Expected Loss= Single Loss Expectation * Annual Occurrence Rate* [19].
- Security Cost-Benefit Analysis: develop a quantitative analysis to calculate the potential business benefit and costs involved with addressing security risks.
- Net Benefit Calculation provides an efficient tradeoff measure: Return Benefit = Annual Expected Loss – Annual Cost of Action [19].
- A business continuity plan (BCP) is needed for each organization.
- Determine the category for tolerable downtimes for the organization's services: e.g. <12 hours, 24 hours, 72 hours, 7 days, 30 days.
- Develop an Information-Security Policy: a policy document is needed that describes what is and what is not permissible use of information in the organization and the consequences for violating the policy.
- The Policy document includes: access-control, external-access, user and physical policies.

- The Policy should be developed by a policy-committee with members from user-groups and stakeholders.
- The policy-committee should meet regularly and should be updated with the organization's needs and current laws.
- Good training and communication of a new policy is needed for awareness.

Further reading with best practices using summarized guidelines for organizations can be found in [10].

IV. Essence of Information Security Standards and Information Security Policies

The necessity for developing and conforming to IT and information security standards at the business or institutional level cannot be understated or emphasized enough since it provides a multi-layer protective shield to many of the security deficiencies and consequent vulnerabilities described in this paper. One example of a set of standards considered as highly relevant to the domain of IT and information processing is that of the WWW Consortium (W3C) Web standards, which are developed with the aim of attaining two key agendas, namely; (i) design principles that includes; Web-for-All (human communication, commerce and knowledge-sharing available to all people, hardware types, software, network infrastructures, native languages, geographic locations, and physical/mental abilities) and Web-on-Everything (all types of web-access devices) , and (ii) a Vision for W3C standards that includes; Web-for-Rich-Interaction, Web-of-Data-and-Services, and a Web-of-Trust [20]. Effectively, such 'web-standards' have established technologies for creating and interpreting web-based content designed to benefit users while remaining compatible with future Web-developments [21].

Other standards particularly relevant to the information-security domain include the ISO 27001 and ISO 27002 standards which establish protocols and guidelines for different levels of security policies within an organization. ISO 27001 formally specifies an Information Security Management System (ISMS) that includes a suite of activities for the management of information security risks and covers all sizes and types of organizations (commercial enterprises, government agencies and non-profit) and industries/markets (retail, healthcare, defense, banking, government and education) [22]. Additionally, the ISO 27001 can be used as the basis for formal compliance assessment by accredited certification bodies in order to certify an organization.

Similarly, the ISO 27002 standard is also relevant to all types of organizations that handles and depends on information processing. This standard explicitly refers to the security of all forms of information, and is not only limited to IT-systems security (e.g. cyber-security). However, whilst the ISO 27001 specifies a mandatory for implementing an ISMS, the ISO 27002 standard specifies suitable controls within the ISMS and is presented as a Code-of-Practice complementary to the ISO 27001 standard [23]. Furthermore, organizations cannot obtain certification by an accredited body through adherence to the ISO 27002. Hence, ISO 27002 is a standard which is normally used more flexibly in accordance to an organization's context [23]. In short, every organization should develop its own information-security policy based on a standard (e.g. such as ISO 27001 with/without ISO 27002). An example document-structure for an organizational policy is provided in [24]. Once a policy-document has been developed, some training for IT staff and employees is required to ensure all are clear of what is required at all levels of responsibility.

V. Conclusions and Open Research Issues

The rapid growth of the Internet and the World Wide Web (WWW) suggests that more attention is required for the security and protection of online sensitive data at various levels. There is an essence and need for Information Assurance in the digital community that encompasses the protection of information in the public and private sectors, academia, or other purposes. Those various sectors are required to take the necessary technical and administrative measures to protect its information assets. In this paper, a number of remarkable data-breach cases and their trends and statistics in the IT sector were shown, along with the technical and organizational-techniques for mitigating such attacks.

Emerging challenges and open research issues which persist in the domain of information security includes: mobile-security, scripting-languages and web-security, and cloud-based security. A notable trend for the development of a more complete information security approach was observed in the literature and related products in the marketplace, which includes: the encryption of something you have or wear (e.g. personal smart-phones) and the encryption of what you are (e.g. using biometric-data). Finally, Figure 2 summarizes and classifies the future research directions and open-issues as a result of our analysis and research findings from a number of recent works.

References

[1] O. Tayan, M.N. Kabir, Y.M. Alginahi, "A Hybrid Digital-Signature and Zero-Watermarking Approach for Authentication and Protection of Sensitive Electronic Documents", The Scientific World Journal, Hindawi Publishing Corporation, Volume 2014, Aug 2014.

[2] O. Tayan, Y.M. Alginahi, "A Review of Recent Advances on Multimedia Watermarking Security and Design Implications for Digital Quran Computing", International Symposium on Biometrics and Security Technologies, ISBAST'14, August 2014

[3] L. Laouamer, O.Tayan, "A Semi-Blind DCT-Watermarking Approach for Sensitive Text-Images", Arabian Journal for Science and Engineering, Mar. 2015.

[4] J. Aguilà Vila, J. Serna-Olvera, L. Fernández, M. Medina and A. Sfakianakis, "A professional view on ebanking authentication: Challenges and recommendations" *9th International Conference on Information Assurance and Security (IAS),* Gammarth, 2013, pp. 43-48.

[5] A. Gawanmeh, H. Al-Hamadi, M. Al-Qutayri, Shiu-Kai Chin and K. Saleem, "Reliability analysis of healthcare information systems: State of the art and future directions" *17th International Conference on E-health Networking, Application & Services (HealthCom)*, Boston, 2015, pp. 68-74.

[6] W. C. Hsieh, C. C. Wu and Y. W. Kao, "A study of android malware detection technology evolution" *Security Technology (ICCST), 2015 International Carnahan Conference on*, Taipei, 2015, pp. 135-140.

[7] Y. Zou, J. Zhu, X. Wang and L. Hanzo, "A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends" in *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1727-1765, Sept. 2016.

[8] A. Sajid, H. Abbas and K. Saleem, "Cloud-Assisted IoT-Based SCADA Systems Security: A Review of the State of the Art and Future Challenges" in *IEEE Access*, vol. 4, no. , pp. 1375-1384, 2016.

[9] L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information Security in Big Data: Privacy and Data Mining" in *IEEE Access*, vol. 2, no. , pp. 1149-1176, 2014.

[10] Symantec Labs, “Internet Security Threat Report”, Technical Report, Published Online, Vol. 21, April 2016.

[11] McAfee Labs “Threats Report”, Online Technical Report, March 2016.

[12] Bruce Snell, “Mobile Threat Report – What’s on the Horizon for 2016”, McAfee Labs, Technical Report, 2016.

[13] http://www.go-gulf.com/blog/cyber-crime/

[14] http://www.infosec.gov.hk/english/business/other_sywa_1.html

[15] http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

[16] http://security.uconn.edu/

[17] http://www.cio.com/article/2384809/e-commerce/15-ways-to-protect-your-ecommerce-site-from-hacking-and-fraud.html

[18] http://www.creativebloq.com/web-design/website-security-tips-protect-your-site-7122853

[19] C.V. Brown, D.W. Dehayes, J.A. Hoffer, E.W Martin, W.C. Perkins, “Managing Information Technology”, Prentice Hall, 7th Edition, 2012.

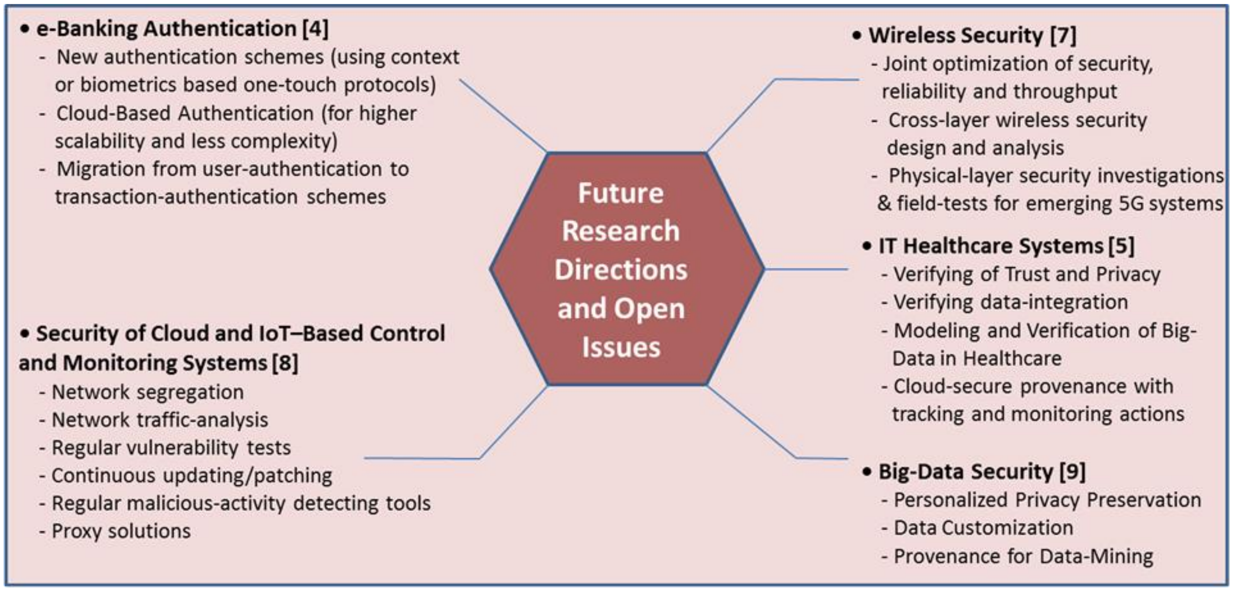
[20] World Wide Web Consortium, http://www.w3.org/Consortium/

[21] Web Standards Mission, http://www.webstandards.org/about/mission/

[22] ISO27001 standard, http://www.iso27001security.com/html/27001.html

[23] Introduction to the ISO27002 standard, http://www.iso27001security.com/html/27002.html#Section6

[24] http://www.computerweekly.com/feature/How-to-create-a-good-information-security-policy



- **e-Banking Authentication [4]**
  - New authentication schemes (using context or biometrics based one-touch protocols)
  - Cloud-Based Authentication (for higher scalability and less complexity)
  - Migration from user-authentication to transaction-authentication schemes
- **Security of Cloud and IoT–Based Control and Monitoring Systems [8]**
  - Network segregation
  - Network traffic-analysis
  - Regular vulnerability tests
  - Continuous updating/patching
  - Regular malicious-activity detecting tools
  - Proxy solutions

Future Research Directions and Open Issues

- **Wireless Security [7]**
  - Joint optimization of security, reliability and throughput
  - Cross-layer wireless security design and analysis
  - Physical-layer security investigations & field-tests for emerging 5G systems
- **IT Healthcare Systems [5]**
  - Verifying of Trust and Privacy
  - Verifying data-integration
  - Modeling and Verification of Big-Data in Healthcare
  - Cloud-secure provenance with tracking and monitoring actions
- **Big-Data Security [9]**
  - Personalized Privacy Preservation
  - Data Customization
  - Provenance for Data-Mining

Fig. 2. Summary and Classification of Future Directions and Open Issues in Information Security

Security and Privacy Risks Awareness for Bring Your Own Device (BYOD) Paradigm

Manmeet Mahinderjit Singh, Chen Wai Chan and Zakiah Zulkefli
School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

***Abstract*—The growing trend of BYOD in the higher education institutions creates a new form of student learning pedagogy in which students are able to use the mobile devices for their academic purposes in anywhere and anytime. Security threat in the paradigm of BYOD creates a great opportunity for hackers or attackers to find new attacks or vulnerabilities that could possibly exploit the students' mobile devices and gains valuable data from them. A survey was conducted in learning the current awareness of security and privacy importance in BYOD for higher education in Malaysia. Based on the analysis of this survey, it demonstrates that the trend of BYOD in Malaysia has begun. Thoroughly, the survey results have been proven that the current basic fundamental security and privacy awareness and knowledge on mobile devices or applications is important in order to protect their mobile devices or data.**

***Keywords*—Mobile Computing; BYOD Higher Education; Security; Privacy; Malicious Software; Risk**

I. Introduction

A tremendous growth in lightweight, portable computing devices and wireless communications has made mobile computing becomes a next-generation computer technology that would transform the way of people interact with each other not just in their daily life, even in their student life. Smartphones, tablets, PDAs, USB memory drives are the perfect examples of mobile computing devices. These devices are equipped with services such as file transfer, internet browsing, mailing services and web-based applications, where these services allow the users to access data or information and collaborate with each other on the move [1]. Thus, there is an increasing use of mobile computing devices among students for their higher education purposes as these devices provide an opportunity for better productivity, performance, convenience and also a promise of mobility [1]. This phenomena leads to a trend which is known as Bring Your Own Device (BYOD) Higher Education, where students are allowed to bring their own devices into their study place and they can perform their educational activities through their own devices.

The increasing use of mobile devices or apps among the students and the growing popularity of the BYOD Higher Education have led to a serious security and privacy attacks towards the campus data and network as well as student's personal information stored on their devices such as student's records, grades, financial and research information and etc. For example, spear phishing [2], Advanced Persistent Threat (APT) attack [3, 4] and malware [4] are the potential attack vectors for BYOD models. Other than that, most of the educational institutions have allowed some form of BYOD trend onto their campus via Network Access Control (NAC) without implementing any organized BYOD policies. This approach poses risk towards the institution's networks as well as student devices such as unauthorized access, attacks of malware and viruses, loss of data and etc. [5].

Thus, the main aim of this study is to observe the awareness of BYOD paradigm and the security and privacy threats that occur within this environment. The objectives of the survey are 1) to investigate the growing trend of BYOD Higher Education and 2) to examine the student's security and privacy knowledge and awareness on mobile devices or applications. The significance of the survey is to understand the perception of the current generation Y on the concept of BYOD and its security challenges that comes with it.

The remainder of the paper is structured as the following. Section 2 covers the background study of BYOD and its security and privacy issue in higher education. Section 3 focusing on BYOD survey conducted to increase the user awareness. Section 4 and 5 are respectively provided a discussion on the survey and the conclusion thoroughly.

II. Background and Related Works

This chapter comprised of three parts included BYOD in higher education, m-learning adoption and security and privacy risks on BYOD.

A. BYOD in Higher Education

Bring Your Own Device (BYOD) Higher Education refers to the practice of higher education students using their own mobile computing devices in their lecture hall or classrooms. Traditionally, BYOD devices include laptops, PDAs, but recently the growing usage of smartphones and tablets have become a part of BYOD devices which offers a high degree of mobility and flexibility. Yu [6] observed that there are three major ways where smartphones were being used in higher education, such as using inbuilt web browsers to access educational materials through online, and using mobile applications to access and interact with a course or lecture content. There also have other uses of smartphones such as recording lectures and seminars, participating in-class polls, logging academic data, taking notes, scanning documents and etc. Tablets such as Apple iPad, Amazon's Kindle, and Samsung Galaxy Tab offer students and lecturers a portable tool by expanding the connectivity and mobility of smartphones through a larger screen and processing power. Johnson, et al. [7] have defined the uses of tablets in education with this statement: "… have gained traction in education

This work is supported by the Fundamental Research Grant Scheme (FRGS) of Universiti Sains Malaysia.

because users can seamlessly load sets of apps and content of their choosing, making the tablet itself a portable personalized learning environment".

B. M-Learning Adoption

According to Akour [8], there are various important factors that influence the acceptance of mobile learning among university students, such as student readiness, ease of access, quality of services, extrinsic influences and institutional commitment. In Lippincott [9], it's believed that the increased capabilities of mobile devices could lead to a new form of engagement with student learning pedagogy. The author also states her belief that student use of mobile devices in higher education correlates to their major area of study as well.

Nowadays, mobile services are provided widely in many universities, for example Massachusetts Institute of Technology (MIT), Stanford University, Harvard University and etc. Besides, the use of mobile learning (m-learning) services in higher education are also becoming one of the active topics in research [10]. In Alzaza and Yaakub [11], the authors explain that the concept of m-learning is considered as the next form of e-learning using mobile technologies so that teachers and learners are able to conduct their learning process in anywhere and anytime. In addition, the authors also conducted a study on the students' awareness and requirements of mobile learning services among Malaysian students in the higher education environment. The results indicate that the students have adequate knowledge and awareness to use the mobile technologies as their choice of the learning environment. M-learning also provides various advantages including freedom to study, low cost, timely application [12], authentic and reliable learning situations, ease of use, support in learning situations [13] , fast production of digital learning materials and flexibility of learning [14].

M-learning could provide lecturers to encourage students to use their devices and some collaborative tools that they support in order to work together on assignments in both physical and virtual learning environments [15]. Other than that, some of the mobile applications could increase the interaction between students and lecturers through an in-class tasks, where it allows students to learn in their preferred places and foster a student-centered learning approach [16].

C. Security and Privacy Risks on BYOD

In this section, security and privacy list will be discussed thoroughly.

1) Installation of malicious software on BYOD

According to Bandara, Ioras, and Maher [17], university and college students are the biggest user of social media or social networking apps such as Facebook, Twitter and YouTube. However, this can lead malwares and viruses such as Wildfire, hosting and spreading throughout the student's personal devices. Bradley, Loucks, Macaulay, Medcalf, and Buckalew [18] explained that an accidental malware downloads not only infects the device itself, it also can easily spread to an entire organization's network within a few seconds. A common mobile malware attack such as Dream Droid [19] and DroidKungfu [20] are luring users to click on malicious web links from their smartphones' web client and install malicious payloads. Moreover, BYOD can easily targeted by hackers to break into someone's devices by sending malicious software through email or application download [21] , for example, hackers always lookout for an opportunity to fool the students through using email or web accounts to spoof the official school mailings as well as bank accounts.

Therefore, students may be encountered the risk of being the victim of a phishing scam that will result in malware or ransomware downloads [22]. As a result, once the student has download and execute malware, the possibility of leakage of student's personal information will increase and the ability of the attacker to steal sensitive information by installing backdoor on the campus network will become possible.

2) Use of untrusted mobile OS and applications

With a "hacker" culture has arisen among young adults, especially when the majority of them are considered as "tech-savvy" nowadays, some of the students used to play with their mobile devices and they able to disable the native OS security feature through the techniques which commonly known as "jailbreaking" or "rooting". By jailbreaking or rooting their mobile devices, they allow installing or upgrading their mobile OS and applications for free that are restricted by default [23]. However, jailbreaking or rooting enables unauthorized programs to be installed on mobile devices, which could probably introduce malware into their devices. This might cause students' devices to be compromised as well as the campus network if student connects his or her devices to the campus network. Besides, some students intend to bypass the institution's proxy servers and access to blocked sites through the mobile VPN and some of them may install some applications such as games, entertainment apps, and P2P video streaming apps which has been restricted by the university. In this instance, it potentially opens up a security threat to the campus network as well as the student's personal data [24].

3) Use of untrusted networks

According to Paullet and Pinchot [25], the mobile devices can be used on both secure and unsecure environments. When users connect their devices to an unsecure network such as public Wi-Fi, the devices will open for a variety of security and privacy attacks such as Wi-Fi hijacking, Bluejacking and etc. For example, Wi-Fi hijacking occurs when a hacker is able to intercept the communications between smartphones and unsecured Wi-Fi hotspots, which allows hackers to gain access to someone's usernames, passwords if a user logs in to certain mobile apps or web site. Therefore, students should be alarmed of the potential risk as many mobile devices have settings to allow the device to automatically connect to available rogue wireless access point that might controlled by hackers [26]. Other than that, forward emails to public Web mail services over the cellular network via BYOD, synchronize academic documents using public cloud-based storage services like Dropbox, iCloud and Google Docs, and interact smartphones through voice in the public place may lead to sensitive data leakage [27].

4) *Lack of physical security controls*

Since the BYOD adoption in the higher education allows students to bring their personally-owned devices into their study place, there is a risk of loss of student personal data will be occurred due to BYOD devices are easy to lose or steal within the campus or in the public. Furthermore, it also places an additional risk to the educational institutions as BYOD devices could contain student's credentials to access sensitive institution's resources or data and also could lead to these data being compromised or malicious activities executed via mobile devices that could gather or corrupt data as well [28]. There's also a possibility that an attacker could configure another device to be a duplicate of a legitimate device which appeared to be authentic to the campus network and steal all the information once the cloned device able to access into the network [23].

5) *Privacy risks on BYOD*

In term of BYOD, the privacy aspect always refers to the concerns that the private data such as personal emails, photos, videos, bank statements, social security numbers, chat histories, usernames, passwords and other credentials are exposed to outsiders. While in the context of BYOD Higher Education, sensitive data such as student's personal details and communications, confidential information about students, assessment data, and confidential institutional data and even the personal credentials for certain educational mobile apps or social networking apps could be exposed to various privacy issues. In Ismail et al. [29], the author states that students are concerned about their privacy and security when they are using m-learning or educational applications, they are worried that their confidential information such as assessment results might be revealed to others.

Besides, the issues of mobile bullying or cyber bullying exist within the educational institutions as the mobile devices can be used for bullying other students and teachers. Some mobile bullying examples like photographing or videotaping other students or teachers and publicly posting, sending or forwarding their photos and videos on the social networking site, websites, emails or message boards in order to harass or humiliate them. Some people even get access and copy or delete someone's information like electronically submitted or stored assignments and homework, or important emails. In addition, some bullies also try to impersonate or pretend to be someone where his or her account has been hacked in order to send abusive calls, texts or images through a mobile phone. This would properly pose a privacy threat to students and teachers personal mobile data.

Other than that, the device location tracking issues also one of the serious privacy issues for the context of BYOD. Although location tracking via a mobile device's geolocation service or GPS are useful for locating lost devices, but illegitimate tracking can cause a serious privacy concern for mobile users. Mobile device tracking or location snooping may expose a threat for students as their location has been recorded and potential criminals will spy on the targeted student's daily activities and perform their crimes. Nowadays, many legitimate or third-party mobile apps provide not just the capability of device tracking, it also allows the tracking of mobile usage behavior tracking via the installed application. This means that the installed apps allow the tracking of selected events occurred on mobile devices and recording every action taken by mobile users [23]. This probably poses another privacy threat to the student if they installed a variety of third party apps.

III. Survey on Security and Privacy Awareness for BYOD Higher Education

The aims of designing this survey are:

- To investigate the growing trend of BYOD Higher Education.
- To examine the student's security and privacy knowledge and awareness on mobile devices or applications.

Fig. 1 provides an overview of the survey samples, methods and variables used in this research work:

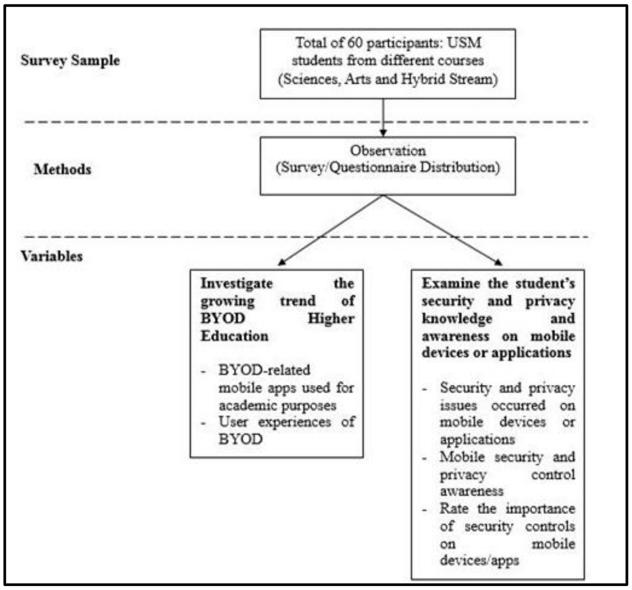

Survey Sample
Total of 60 participants: USM students from different courses (Sciences, Arts and Hybrid Stream)
Methods
Observation (Survey/Questionnaire Distribution)
Variables
Investigate the growing trend of BYOD Higher Education
- BYOD-related mobile apps used for academic purposes
- User experiences of BYOD
Examine the student's security and privacy knowledge and awareness on mobile devices or applications
- Security and privacy issues occurred on mobile devices or applications
- Mobile security and privacy control awareness
- Rate the importance of security controls on mobile devices/apps

Fig. 1. Survey sample, variable and methods

In this research, a data collection method has adopted: Observation. Observation is the method that the compilation of data collected through questionnaires or survey. For example, an online survey will be generated, conducted, distributed and get responses through online. As a result, a total of 60 university students of different courses from the University Sains Malaysia (USM) have taken part in this survey regarding the security and privacy of BYOD Higher Education, where 39 of them are males and 21 of them are females.

IV. Survey Results and Findings

In this section, the survey findings will be displayed.

A. Investigate the Growing Trend of BYOD Higher Education

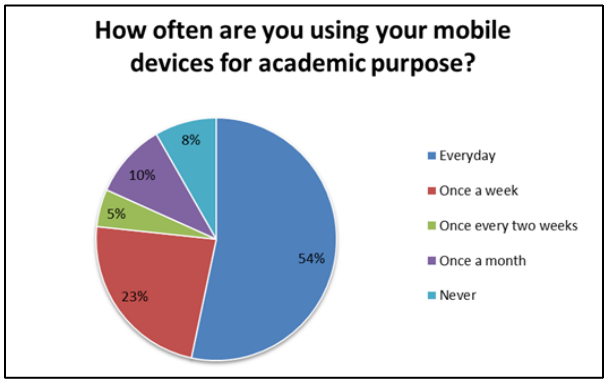


Fig. 2. Use of mobile devices for academic purposes

This survey examines the use of mobile devices for academic purposes among the university students as well as the adoption of a BYOD trend throughout the higher education. A single choice closed question will be asked about how often the participants on using mobile devices for academic purposes. Based on the survey result as shown in Fig. 2, there are 32 participants (54%) use their mobile devices for academic purposes every day, 14 participants (23%) use their mobile devices once a week, 3 participants (5%) use their mobile devices once every two weeks, 6 participants (10%) use their mobile devices once a month and 5 participants (8%) not using their mobile devices for academic purposes.

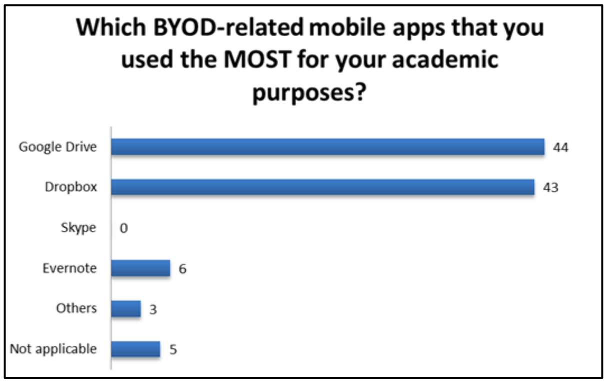


Fig. 3. Use of BYOD-related mobile apps for academic purposes

Next, a multiple choice question will be asked on which BYOD-related mobile apps that the participants used the MOST for their academic purposes. Based on the survey result from Fig. 3, there are two BYOD-related mobile apps that our participants used the most for their academic purposes: Google Drive (73%) and Dropbox (72%), followed by a note taking app, Evernote (10%) and other mobile apps such as OneNote, Notepad, WPS. Apparently, participants are not using Skype mobile app (0%) for their academic purposes. However, there are 5 participants are not using any mobile apps for their academic purposes as well.

B. Examine the Student's Security and Privacy Awareness and Knowledge on BYOD Mobile Device and Apps

In this survey, participants also tested on several questions that are related to the security and privacy awareness on downloading the BYOD-related mobile apps as well as their opinions on how to secure their personal mobile data after installing and using these mobile apps for their learning process. From the survey result shown in the Fig. 4, there are 58% of the participants answered that they are viewing on the reviews of the BYOD-related mobile apps before they download and install them, while the rest of the participants (42%) are not viewing on the reviews before they download and install the apps.

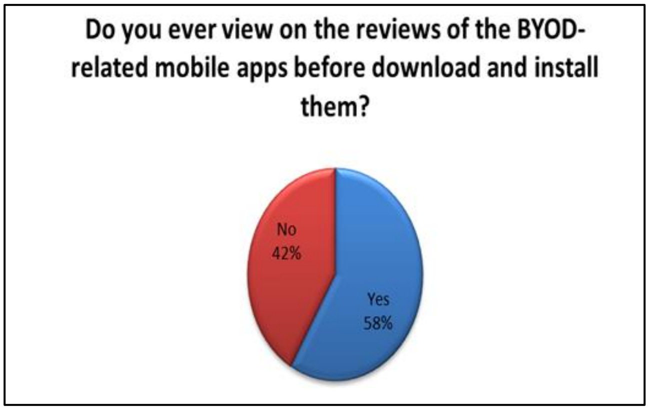


Fig. 4. View on reviews of the BYOD-related mobile apps

From the survey result shown in Fig. 5, there are 89% of participants answered that they have not read and understand the BYOD-related mobile app's "Privacy Policy" before they download and install the app, while there are only 12% of participants answered that they have read and understand the "Privacy Policy" before they download and install.

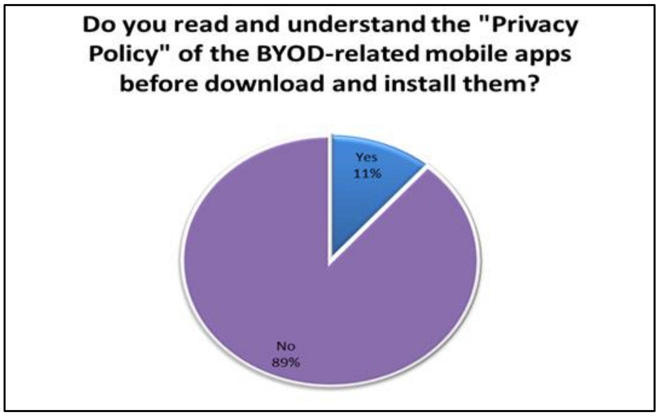


Fig. 5. View on privacy policy of the BYOD-related mobile apps

From the survey result shown in Fig. 6, there are 43% of participants answered that they have checked and read the app permissions that the BYOD-related mobile app could be accessed before they accept and install the app, while 57% of participants answered that they have not read the app permissions before they accept and install the app. So, there are still a lot of mobile users not checking on the app permissions to access sensitive information, such as mobile data, such as identity, contacts, location and device id.

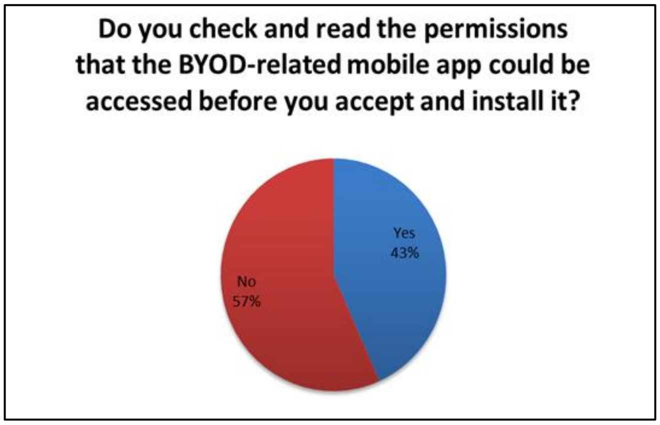


Fig. 6. Check the permissions of the BYOD-related mobile apps

For the next question, the participants are required to provide their opinions on how they feel that is it reasonable for an app to access too much personal data such as identity, contacts, location and device id through the stated app permissions before they decide to install the app. From the survey result shown in Fig. 7, almost 92% of participants answered that it is not reasonable for a mobile app to access too many personal data through the stated app permissions. Whereas, only 8% of participants answered that it is reasonable for an app in order to access these personal data by providing several reasons such as: (a) some mobile apps can be more reliable and useful once the apps accessed these personal data, (b) some mobile apps will automatically setup the app account with syncing the personal data, (c) personal data which accessed by the mobile apps will not easily expose to someone else.

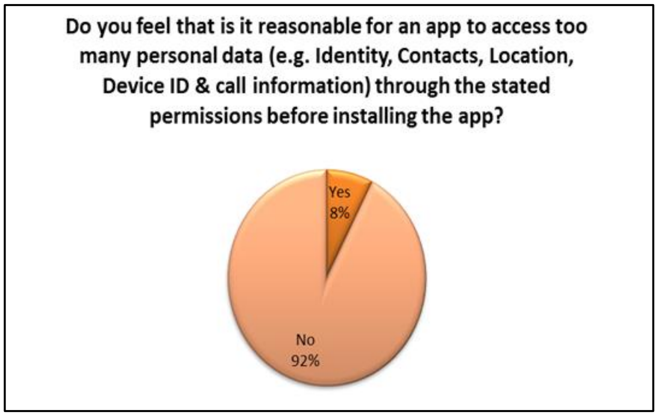


Fig. 7. Opinions on accessibility of mobile apps through permissions

From the survey result shown in Fig. 8, there are 85% of participants answered that they do not feel secure when they are signing in their BYOD-related mobile app's accounts when the app does not have any indication of SSL connection or a visible HTTPS indicator. Whereas, there are 15% of participants answered that they are feeling secure when signing in the BYOD-related mobile app's accounts even though there is no HTTPS indicator.

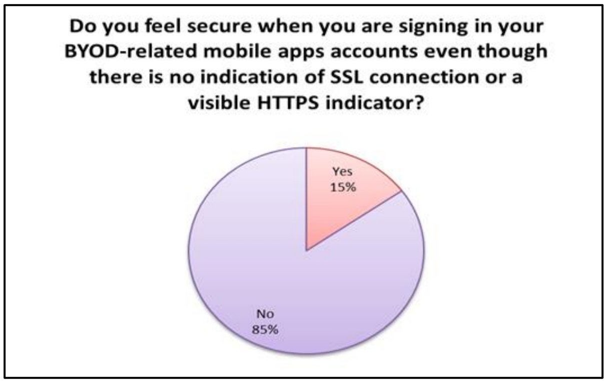


Fig. 8. Opinions on SSL connection for BYOD-related mobile apps

The reasons of some participants answered that they are feeling secure when they sign in their BYOD-related mobile app's accounts even though there is no indication of SSL connection are: (a) some of the participants believed that there have many researchers around the world focus on these mobile app security, (b) some of the participants assume that these mobile apps have a strong security since these mobile apps developed by well-known developers, (c) some participants assume these mobile apps are secured based on the stated policy, (d) mainstream mobile apps from Google or Dropbox are generally secure and there will no security issues should be worry about.

In Fig. 9, there are 83% of participants do not share their BYOD-related mobile app's personal login credential to their family or someone who close to them, whereas there are only 13% of participants are sharing their login credential to their family or people who close to them. However, only 4% of participants did not know their login credentials have been shared with someone else.

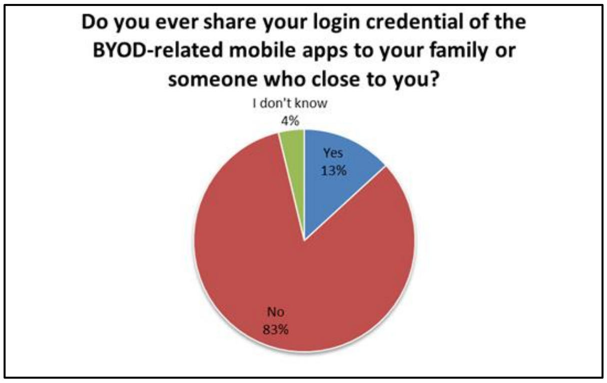


Fig. 9. Sharing login credential of the BYOD-related mobile apps

In Fig 10, nearly 38% of participants did not know the security feature of two-step verification for mobile apps, and 36% of participants did not use two-step verification for their BYOD-related mobile apps. On the other hand, 26% of participants are using two-step verification to verify their BYOD-related mobile apps.

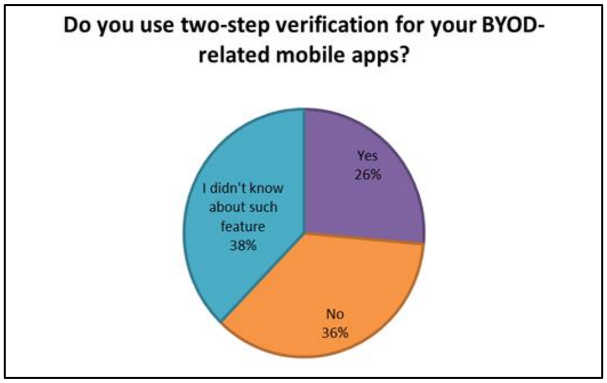


Fig. 10. Use of two-step verification for BYOD-related mobile apps

For the next question, participants will be asked for the opinion on how to create a strong and complex password for the login credentials of mobile apps. Most of the participants have provided similar answers, which is a strong and complex password can be created by the combination of capital and small letters, mixture of alphabetic, numeric and special characters/symbols, password length should be at least 8 characters, passwords should not relate to any personal information such as phone number, IC number and date of birth. Besides that, different accounts should have different passwords. Some of the participants also recommended the implementation of both password and biometric credentials.

In Fig. 11, there are 60% of participants are not syncing or sharing their personal files through BYOD-related mobile apps while accessing the public network or open area Wi-Fi. On the other hand, 40% of participants are accessing the open area Wi-Fi from any locations (ie: coffee shops and airport) when they are syncing their BYOD-related mobile apps' files.

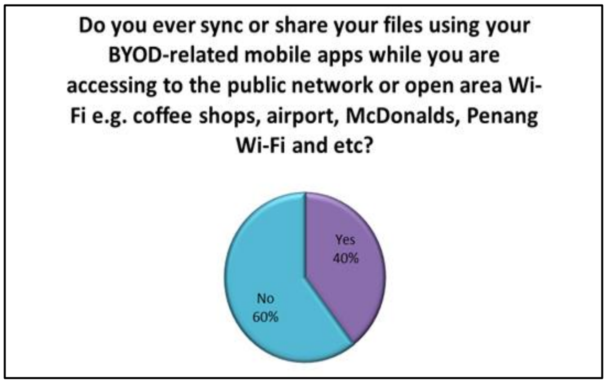


Fig. 11. Access BYOD-related mobile apps through public Wi-Fi

From the survey result shown in Fig. 12, there are nearly 60% of participants are not setting up any passcode/pin-code or add an encryption option in order to keep their personal files in a mobile encrypted and safe. Whereas, there are 34% of participants are adding passcode or encryption option to protect their valuable files on their mobile devices. In addition, there also have 6% of participants did not know the security feature of passcode or encryption option.

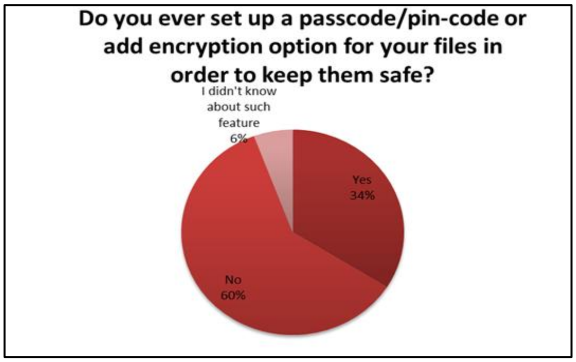


Fig. 12. Set passcode/encryption options for mobile files

Next, we will let the participants proceed to the questions related to the rating of several security and privacy concerns that will impact the mobile devices. For the security concerns, the research has listed out several concerns, such as:

- Lost or stolen mobile devices with personal data
- Malicious applications downloaded to the mobile devices
- Connected mobile devices to unsecured Wi-Fi or networks
- Accessed to insecure web browsing
- Lack of security patches on mobile apps
- High rate of users changing or upgrading their mobile devices
- Lack of efficient encryption methods on mobile apps

In the Fig. 13, for the first security concern: Lost or stolen mobile devices with personal data, where there are 33 participants or more than 50% of participants rated this security concern that can provide 100% of the impact of the security of mobile devices. For the second security concern: Malicious applications downloaded to the mobile devices, where most of the participants are rating this security concern that provides 80-100% of the impact (40 participants in total) to the security of mobile devices. Whereas for the third security concern: Connected mobile devices to unsecured Wi-Fi or networks, where the majority of the participants (20 participants) rated this security concern can provide 50% of the impact of the mobile devices.

For the fourth security concern: Accessed to insecure web browsing, where there are 19 participants (32%) rated it as it can provide 50% of the impact of the mobile devices instead of 100% of the impact of the mobile devices. The fifth security concern: Lack of security patches on mobile apps is quite similar to the fourth security concern, where there are 23 participants (38%) rated it as it can provide 50% impact of the mobile devices. On the sixth security concern: High rate of users changing or upgrading their mobile devices, where the majority of the participants (19 participants) have rated this security concern as it can provide 50% of the impact of the mobile devices. For the last security concern: Lack of efficient encryption methods on mobile apps, where the majority of the participants (24 participants) also rated this security concern as it can provide 50% of the security impact of the mobile devices.

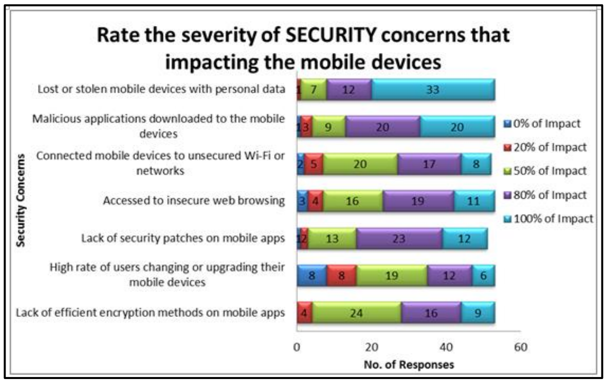


Fig. 13. Severity of security concerns that impacting mobile devices

For the privacy concerns, the research has listed out several concerns such as:

- Mobile data/information leakage
- Mobile user's location has been tracked and collected
- Collection on user identity by service providers
- Lack of transparency on mobile app permission
- Sniffing and snooping on mobile phone sensors
- Losing access of ownership of mobile data

In Fig. 14, for privacy concern, such as mobile data/information leakage, there are 29 participants rated it as it can provide 100% of the impact to the mobile devices. For the second privacy concern: Mobile user's location has been tracked and collected, there are also have 25 participants rated it, as it can provide 100% of the impact of the mobile devices. For the third privacy concern: Collection on user identity by service providers, there are 22 participants rated this privacy concern can provide 80% of the impact of the mobile devices rather than it can provide 100% of the impact of the mobile devices.

Next, for the fourth privacy concern: Lack of transparency on mobile app permission, there have a majority of 20 participants rated it as it can provide 80% of the impact rather than 100% of the impact to the mobile devices. For the fifth privacy concern: Sniffing and snooping on mobile device sensors, there has a majority of 22 participants rated this privacy concern as it can provide 80% of the impact of the mobile devices. For the last privacy concern: Losing access of ownership on mobile data, there has a majority of 21 participants rated it as it can provide 100% of the impact of the mobile devices.

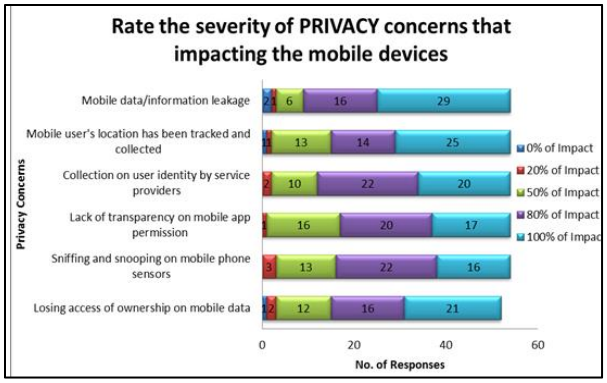


Fig. 14. Severity of privacy concerns that impacting mobile devices

In order to improve the student awareness on the security and privacy issues inherent on different types of BYOD-related mobile apps, there are a set of defined security controls such as: Authentication, Authorization, Confidentiality, Integrity, Availability and Non-Repudiation are taken into account to determine which security controls will be the most important for protecting the mobile data.

In Fig. 15, there are 67% of participants rated that Authentication is the most important security control for mobile user to protect their mobile data. Whereas, there are 23% of participants also indicated that Authentication is a very important security control which is needed for mobile devices.

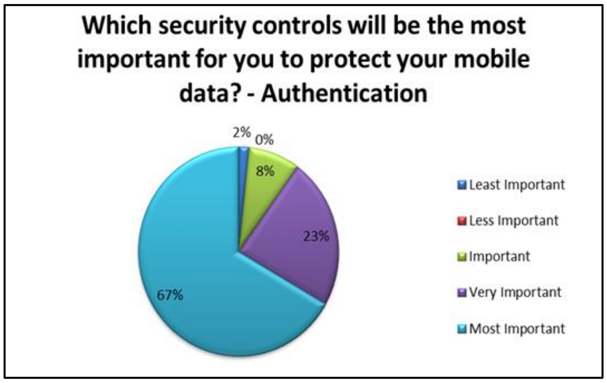


Fig. 15. Security controls for mobile data: Authentication

In Fig. 16, there are 52% of participants rated that Authorization is one of the most important security controls for mobile devices followed by 35% of participants also rated it as it is very important.

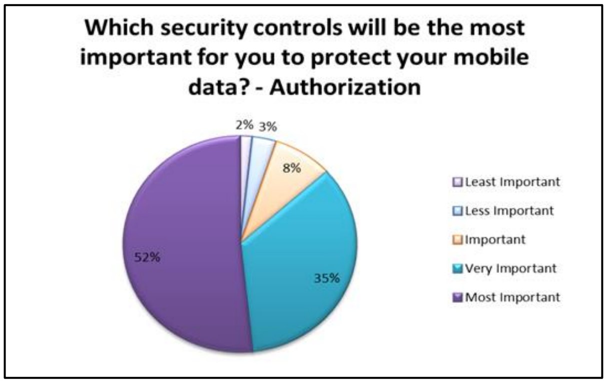


Fig. 16. Security controls for mobile data: Authorization

Similar to previous two security controls, there are more than half of participants rated the security control: Confidentiality (as shown in Fig. 17) as the most important security control for mobile devices followed by 30% of participants rated it as a very important security control.

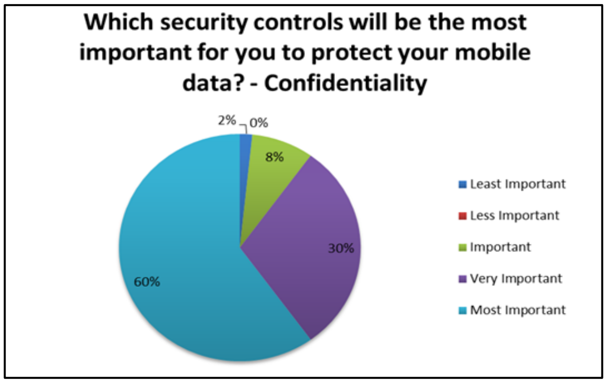


Fig. 17. Security controls for mobile data: Confidentiality

For Integrity security control (see Fig. 18), there also have 62% of participants rated this security control as the most important security control for protecting the mobile data and 30% of participants rated it as very important.

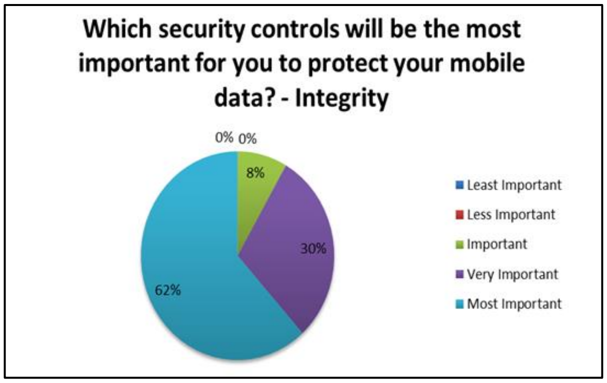


Fig. 18. Security controls for mobile data: Integrity

Unlike the previous security controls, there have only 43% of participants rated Availability (see Fig. 19) as the most important security control of mobile devices while 37% of participants rated it as very important and 20% of participants rated it as important.

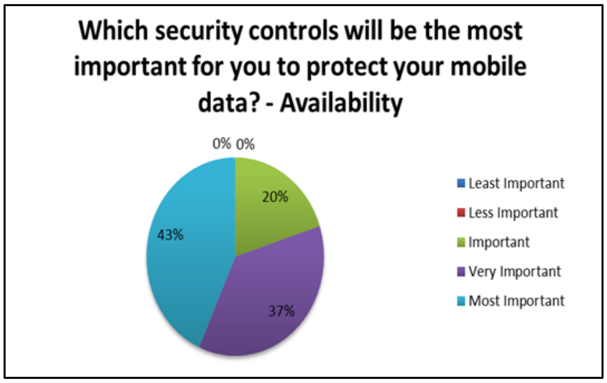


Fig. 19. Security controls for mobile data: Availability

For the last security control: Non-repudiation (see Fig. 20), there have about 55% of participants rated it as the most important security control of mobile devices and 30% of participants rated it as very important and the rest of participants rated it as important.

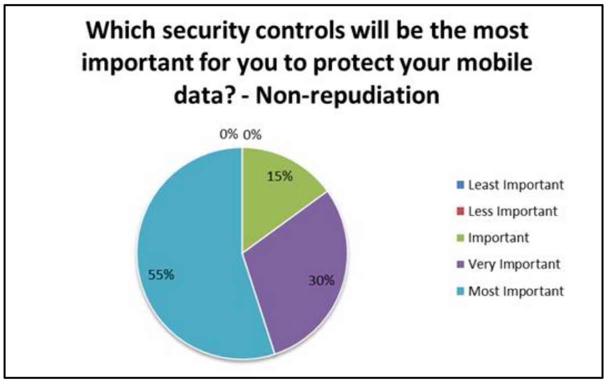


Fig. 20. Security controls for mobile data: Non-Repudiation

V. Survey Discussions

Survey discussion is given in this section.

A. *Growing Trend of BYOD Higher Education*

Based on the survey results stated from Fig. 2 and Fig. 3, it could be concluded that the trend of BYOD is starting to adopt into the higher education system since there are quite a number of USM students started to use their own mobile devices for their academic purposes instead of their personal usage. Statistics have been proven that many USM students started to use some of the popular mobile applications such as Google Drive, Dropbox as their educational apps in order to use these apps throughout their student life in university.

B. *Security and Privacy Awareness on BYOD Mobile Apps*

In this survey, the participants are being tested on several questions that are related to the security and privacy awareness when downloading the BYOD-related mobile apps. Based on the survey results obtained from Fig. 4, it can be proven that there are quite a number of participants that are still considering the security and safety of downloading and installing certain mobile apps that might be harmful for their mobile devices and their personal mobile data by determining the positivity or negativity of the user reviews posted on

Google Play Store or other sources. But, Fig. 5 and Fig. 6 show that most participants are often just skipped the security precautions such as checking and reading “Privacy Policy” and app permissions before they want to download and install the mobile apps. This probably came out an assumption that most of the mobile users do not really want to spend their time on reading and understanding the policies and app permissions that written by the app developers in order to notify mobile users that what kind of data that could be collected by the app itself.

Next, Fig. 8 shows that there are quite a number of participants (85%) are aware the importance of SSL connection or a visible HTTPS indicator for signing in the mobile app’s accounts, which probably suggest that many participants take the security of their mobile app’s account seriously. Fig. 11 shows that most of the participants are not syncing or sharing their files stored in BYOD-related mobile apps while accessing to the public network or Wi-Fi, this indicates that many participants aware of the danger of sharing or sync any mobile data through public Wi-Fi.

Besides, Fig. 10 and Fig. 12 show that most of the participants are still did not know how to configure several security features such as two-step verification, pin codes and encryption option in order to protect their mobile data and applications. It could be concluded that mobile user education and awareness programs are needed for educating the mobile users on how to configure several basic security features for their mobile devices or applications. In addition, Fig. 9 indicates that there has a very high percentage of participants are not sharing their login credentials as well.

Based on the survey results from Fig. 13 and Fig. 14, many participants realized there are several security and privacy concerns that will probably cause a serious impact to their mobile devices. Some of these results are supported by the study of Obodoeze, et al. [30] which demonstrated the various forms of challenging security concerns, including losses of mobile devices, virus and malware attacks and etc., and also the study of Kambourakis [26] that discussed the security and the privacy challenge of m-learning. Furthermore, survey results from Fig. 15 to Fig. 20 show that most of the participants also believed that there should be a list of defined security controls such as Authentication, Authorization, Confidentiality, Integrity, Availability and Non-repudiation used to enhance the security and privacy strengths of the mobile devices or applications.

As a conclusion of this survey, it has been proven that most of the USM students have a fundamental knowledge or awareness about the security and privacy of the mobile devices or applications, which probably indicates that many students nowadays start concerning the security and privacy risks that could be happening on their mobile devices. The survey conducted here with a small focus group of students in one of the higher educational centers in Malaysia may project results and finding that is closely correlated to the samples size and background. However, to our knowledge, there is yet for a BYOD in higher education research to be done in Malaysia. Thus, the survey results and findings can be generalized as the perception of higher education students or as the viewpoint of the Y generation as a whole.

VI. Conclusion and Future Work

This research start with collecting and studying different approaches on BYOD Higher Education, M-Learning adoption within higher education institutions as well as the security and privacy vulnerabilities and attacks that will be happened on BYOD Higher Education environment in order to find out an effective solution for solving the research problem. Hence, this research moves on to conduct the survey to investigate the security and privacy awareness among university students in BYOD Higher Education. Based on the analysis of this survey, it shows that the trend of BYOD is started among the USM students as many students started to move towards the trend of BYOD where they are using mobile devices or applications for their academic tasks. Besides, the survey results have been proven that the current USM students have a basic or fundamental security and privacy awareness and knowledge on mobile devices or applications where most of the students start concerning the security and privacy controls or services in order to protect their mobile devices or data.

Besides that, the growing trend of BYOD in higher education institutions eventually creates a new form of student learning pedagogy where students able to use the mobile devices for their academic purposes anywhere and anytime. However, this also creates a great opportunity for hackers or attackers to find new attacks or vulnerabilities that could possibly exploit the students’ mobile devices and gains valuable data from them. Hence, a further research in finding new attacks or vulnerabilities on BYOD Higher Education is still necessary in order to increase the security and privacy awareness among the security specialists and university students. Besides, the existing case studies still require some additional research in order to improve the details of the case studies as well as its impacts towards the assets or systems in terms of confidentiality, integrity and availability. This could probably make the case studies more concise, detailed and easy to evaluate using a standardized security metrics framework.

References

[1] B. Alleau and J. Desemery, "Bring your own device: It’s all about employee satisfaction and productivity, not costs!," Capgemini Consulting2013.

[2] L. B. Lau, M. M. Singh, and A. Samsudin, "Trusted System modules for tackling APT via spear-phishing attack in BYOD environment," Undergradute Research Thesis, School of Computer Science, Universiti Sains Malaysia, 2015.

[3] Z. Zulkefli, M. Mahinderjit-Singh, and N. Malim, "Advanced Persistent Threat mitigation using Multi Level Security – Access Control framework," in Computational Science and Its Applications -- ICCSA 2015. vol. 9158, O. Gervasi, B. Murgante, S. Misra, M. L. Gavrilova, A. M. A. C. Rocha, C. Torre, et al., Eds., ed: Springer International Publishing, 2015, pp. 90-105.

[4] M. M. Singh, S. S. Siang, O. Y. San, N. H. A. H. Malim, and A. R. M. Shariff, "Security attacks taxonomy on Bring Your Own Devices (BYOD) Model," International Journal of Mobile Network Communications & Telematics (IJMNCT) vol. 4, pp. 1-17, October 2014 2014.

[5] R. Afreen, "Bring Your Own Device (BYOD) in higher education: Opportunities and challenges," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 3, pp. 233-236, 2014.

[6] F. A. Yu, "Mobile/smart phone use in higher education " in Southwest Decision Sciences Institute Conference, 2012, pp. 831-839.

[7] L. Johnson, S. A. Becker, M. Cummins, V. Estrada, A. Freeman, and H. Ludgate, "The NMC horizon report: 2013 higher education edition," New Media Consortium, Texas2013.

[8] H. Akour, "Determinants of mobile learning acceptance: An empirical investigation in higher education," Ph.D. thesis, Oklahoma State University, 2010.

[9] J. K. Lippincott, "A mobile future for academic libraries," Reference Services Review, vol. 38, pp. 205-213, 2010.

[10] A. Samochadin, D. Raychuk, N. Voinov, D. Ivanchenko, and I. Khmelkov, "MDM based mobile services in universities," in International Conference on Emerging of Networking, Communication and Computing Technologies (ICENCCT 2014) Co-jointed with International Conference on Emerging Trends of Computer Science with Educational Technology (ICETCSET 2014), Zurich, Switzerland, 2014, pp. 35-41.

[11] N. S. Alzaza and A. R. Yaakub, "Students' awareness and requirements of mobile learning services in the higher education environment," American Journal of Economics and Business Administration vol. 3, pp. 95-100, 2011.

[12] N. S. Alzaza and A. N. Zulkifli, "Mobile-Based Library Loan Service (MBLLS)," in Proceedings of the Rural ICT Development Conference ’07 (RICTD’07), Executive Development Centre (EDC), UUM, 2007, pp. 1-8.

[13] P. Seppälä, J. Sariola, and H. Kynäslahti, "Mobile learning in personnel training of university teachers," in Wireless and Mobile Technologies in Education, 2002. Proceedings. IEEE International Workshop on, 2002, pp. 136-139.

[14] M. Sharples, D. Corlett, and O. Westmancott, "The design and implementation of a mobile learning resource," Personal and Ubiquitous Computing, vol. 6, pp. 220-234, 2002.

[15] K.-W. Lai, F. Khaddage, and G. Knezek, "Blending student technology experiences in formal and informal learning," Journal of Computer Assisted Learning, vol. 29, pp. 414-425, 2013.

[16] Holzinger, A. Nischelwitzer, and M. Meisenberger, "Mobile phones as a challenge for m-learning: Examples for mobile interactive learning objects (milos)," in Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on, 2005, pp. 307-311.

[17] Bandara, F. Ioras, and K. Maher, "Cyber security concerns in e-learning education," 7th International Conference of Education, Research and Innovation, pp. 728-734, 2014.

[18] J. Bradley, J. Loucks, J. Macaulay, R. Medcalf, and L. Buckalew, "BYOD: A global perspective harnessing employee-led innovation (survey report)," 2012.

[19] O. Krehel. (2011). Worse than zombies: The mobile botnets are coming. Available: http://www.idt911blog.com/2011/06/worse-than-zombies-the-mobile-botnets-are-coming/

[20] X. Jiang. Security alert: New sophisticated android malware droidkungfu found in alternative chinese app markets. . Available: http://www.csc.ncsu.edu/falculty/jiang/DroidKungFu.html

[21] A., "Information Security Risk Management," 2013.

[22] S. Poremba, "5 higher education information security threats you should know before your child leaves for college," Forbes2014.

[23] D. Caroll, M. Rose, and V. Sritapan, "Mobile Security Reference Architecture," 2013.

[24] M. Potts, "The state of information security," Network Security, vol. 2012, pp. 9-11, 7// 2012.

[25] K. Paullet and J. Pinchot, "Mobile malware: coming to a smartphone near you ?," Issues in Information Systems, vol. 15, pp. 116-123, 2014.

[26] G. Kambourakis, "Security and privacy in m-learning and beyond: Challenges and state-of-the-art," International Journal of U- & E-Service, Science & Technology, vol. 6, pp. 67-84, 2013.

[27] P. Wei, L. Feng, K. J. Han, Z. Xukai, and W. Jie, "T-dominance: Prioritized defense deployment for BYOD security," in Communications and Network Security (CNS), 2013 IEEE Conference on, 2013, pp. 37-45.

[28] Pillay, H. Diaki, E. Nham, S. Senanayake, G. Tan, and S. Deshpande. (2013, Does BYOD increase risks or drive benefits? (Unpublished). Available: http://hdl.handle.net/11343/33345

[29] S. Ismail, S. B. A. Rahman, N. M. Noordin, S. M. S. Mustafa, Z. F. Zamzuri, M. Manaf, et al., "Student perception on security requirement of e-learning services," Procedia - Social and Behavioral Sciences in 6th International Conference on University Learning and Teaching (InCULT 2012), vol. 90, pp. 923-930, 2013/10/10 2013.

[30] F. C. Obodoeze, F. A. Okoye, C. N. Mba, S. C. Asogwa, and F. E. Ozioko, "A holistic mobile security framework for nigeria," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 2, pp. 5-11, 2013.

Improved Mechanism to Prevent Denial of Service Attack in IPv6 Duplicate Address Detection Process

Shafiq Ul Rehman, Selvakumar Manickam
National Advanced IPv6 Centre (NAv6)
University of Science Malaysia
Penang, Malaysia

***Abstract*—From the days of ARPANET, with slightly over two hundred connected hosts involving five organizations to a massive global, always-on network connecting hosts in the billions, the Internet has become as important as the need for electricity and water. Internet Protocol version 4 (IPv4) could not sustain the growth of the Internet. In ensuring the growth is not stunted, a new protocol, i.e. Internet Protocol version 6 (IPv6) was introduced that resolves the addressing issue IPv4 had. In addition, IPv6 was also laden with new features and capabilities. One of them being address auto-configuration. This feature allows hosts to self-configure without the need for additional services. Nevertheless, the design of IPv6 has led to several security shortcomings. Duplicate Address Detection (DAD) process required for auto-configuration is prone to Denial of Service (DoS) attack in which hosts are unable to configure themselves to join the network. Various mechanisms, SeND, SSAS, and the most recent being Trust-ND, have been introduced to address this issue. Although these mechanisms were able to circumvent DoS attack on DAD process, they have introduced various side effects, i.e. complexities and degradation of performance. This paper reviews the shortcomings of these mechanism and proposes a new mechanism, Secure-DAD, that addresses them. The performance comparison between Trust-ND and Secure-ND also showed that Secure-DAD is more promising with improvement in terms of processing time reduction of 45.1% compared to Trust-ND while preventing DoS attack in IPv6 DAD process.**

***Keywords*—Secure-DAD; Duplicate Address Detection; Denial of Service Attack; IPv6 Security; Address auto-configuration**

I. Introduction

Address auto-configuration [1] is the main feature of IPv6 Internet protocol [2]. This mechanism allows IPv6 enabled devices to configure IP addresses automatically without the need for addition services providers such as; DHCPv6, thus provides flexibility in address configuration. However, self-generated IP address has to be unique in order to prevent the conflict of IP address among hosts in IPv6 network [1]. Although, it can be argued that IP conflict is extremely remote due to the immensity of the address space, this will not be the case in the coming years due to the growth in mobile device and new drivers such as; Internet of Things (IoT) [3, 4] and Cloud [4]. Therefore, there is a mechanism known as Duplicate Address Detection (DAD) process [1, 5] to verify the uniqueness of self-generated IP address. In IPv6 network every host must perform DAD process in order to configure a unique valid IP address.

In IPv6 network, for Neighbor Discovery [6, 7] IPv6 hosts use two types of ICMPv6 [8] messages also known as neighbor discovery messages i.e. Neighbor Solicitation (NS) and Neighbor Advertisement (NA) messages. Neighbor solicitation (NS) message is used to send a query to neighboring hosts on same link and in response to that query existing hosts use Neighbor Advertisement (NA) message. While performing DAD process, new hosts send NS message to verify whether the self-generated IP address is already obtained by any existing host on a same link. If any existing host has configured the same IP address then that host replies back with NA message that the self-generated IP address is already configured.

During standard DAD process IPv6 hosts are considered trustworthy. Therefore, IPv6 hosts rely on the information being exchanged on a same link. Thus, malicious host can exploit the DAD process by disrupting the communication during address verification between the hosts. Research [5, 9, 10] have shown that DAD process is vulnerable to denial of service (DoS) attacks. During DoS-on-DAD attack, a malicious host tries to prevent the victim host to configure a unique valid IP address by claiming the existence of self-generated IP address via sending fake NA messages in reply to its NS messages. Hence, victim host is unable to configure its unique IP address. Thus, victim host cannot communicate on a same link due to DAD process failure.

Considering this vulnerability with DAD process, some of the security mechanisms have been proposed such as; SeND [10], SSAS [11], and Trust-ND [12]. SeND mechanism was suggested to solve the security concerns of ND messages. However, this mechanism is not trivial due to its design which possess heavy computation and complexity issues during ND message processing [11, 12]. In order to address this issue, Simple Secure Addressing Scheme (SSAS) was proposed. This mechanism to some extent addressed the issue of complexity by introducing a new scheme compared to the SeND mechanism. However, SSAS still requires significant amount of time to process the ND messages [12]. Recently, Trust-ND has been proposed that claims to be the lightweight mechanism compared to SeND and SSAS schemes. However, the issue with Trust-ND mechanism is that it is built on SHA-1 hashing algorithm which has been found vulnerable to hash collision attacks [13, 14]. Thus, due to its design it can induce DoS attack during DAD process.

This paper introduces a new mechanism know as Secure-DAD which is faster in terms of processing time and effective enough to prevent DoS attack during DAD process. The rest of the paper is organized as follows: Section 2 will present an overview of DAD process and its security issues. Section 3 will discuss the related work. Section 4 explains the design and implementation of Secure-DAD mechanism. Section 5 will present a Test-bed setup environment. Section 6 will discuss the evaluation procedure of Secure-DAD mechanism. Section 7 provides the experimental results and discussion. And finally, Section 8 will present the conclusion and future work.

II. IPv6 DAD Process and its Security Issues

In order to be able to communicate on the same network, IPv6 host(s) has to verify the uniqueness of its self-generated IP address which is the final stage of address auto-configuration [1, 5, 10]. This verification procedure is being executed through Duplicate Address Detection process. New host performs DAD process by sending Neighbor Solicitation (NS) message to all node multicast address (FF02::1) so that existing hosts can receives NS message. NS message carry the tentative IP address that new host has generated and would like to assign it as a preferred address. If that tentative address is configured already by any other host in the network then that particular host will reply back with a Neighbor Advertisement (NA) message. Hence, new host repeats the DAD process again, in case if there is no response to its generated NS message; then it will consider the generated IP address is unique [1, 5, 15]. Thus, a new host can use it as a preferred IP address. Figure 1 describes the DAD process in IPv6 network.

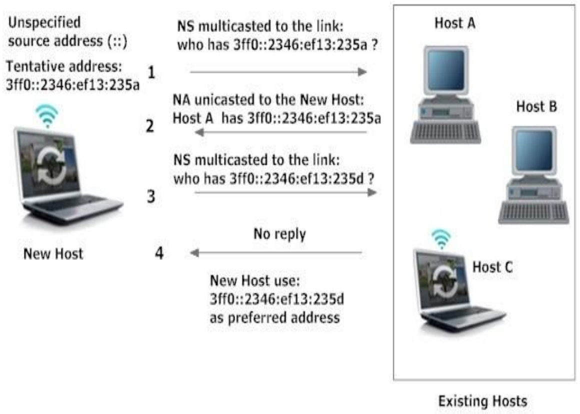

Unspecified
source address (::)
Tentative address:
3ff0::2346:ef13:235a
NS multicasted to the link:
who has 3ff0::2346:ef13:235a ?
1
NA unicasted to the New Host:
Host A has 3ff0::2346:ef13:235a
2
NS multicasted to the link:
who has 3ff0::2346:ef13:235d ?
3
No reply
4
New Host
New Host use:
3ff0::2346:ef13:235d
as preferred address
Host A
Host B
Host C
Existing Hosts

Fig. 1. Duplicate address detection process [15]

In IPv6 link local communication any existing IPv6 host can participate in DAD process. Since, ND messages such as: NS/NA messages are insecure by design. Thus, an attacker can easily exploit the DAD process by fabricating NA message and reply it to every NS message received. This can disrupt DAD process and cause DAD failure. Hence, new host will not be able to obtain a valid IP address. As a result, new host cannot communicate in IPv6 link local network. This attempt of DoS attack is known as DoS-on-DAD attack. Figure 2 illustrates the DoS attack on DAD process in IPv6 network.

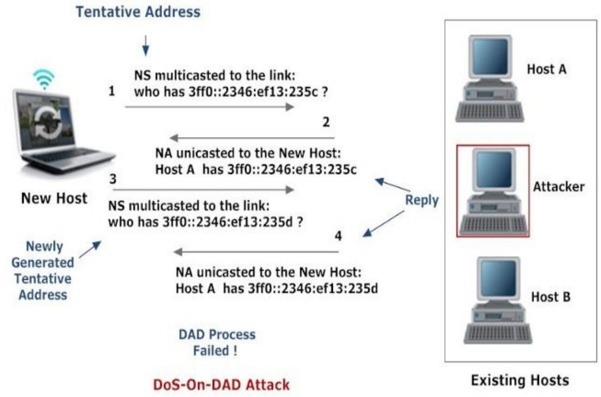

Tentative Address
1
NS multicasted to the link:
who has 3ff0::2346:ef13:235c ?
2
NA unicasted to the New Host:
Host A has 3ff0::2346:ef13:235c
3
NS multicasted to the link:
who has 3ff0::2346:ef13:235d ?
4
NA unicasted to the New Host:
Host A has 3ff0::2346:ef13:235d
New Host
Newly
Generated
Tentative
Address
Reply
Host A
Attacker
Host B
DAD Process
Failed !
DoS-On-DAD Attack
Existing Hosts

Fig. 2. Denial of service attack on DAD process [15]

III. Related Work

Considering the security concern with IPv6 DAD process, existing mechanisms such as: SeND, SSAS, and Trust-ND have been proposed to address this problem in IPv6 link local network. However, these mechanisms have some issues due to their designed mechanism which restrains their implementation on DAD process in IPv6 network. This section describes these issues and limitations with existing mechanisms as follows:

A. *Secure Neighbor Discovery (SeND)*

SeND was introduced to address the security issues related with NDP messages. It introduces four NDP options; CGA option, Nonce option, Timestamp option, and RSA signature option as well as two ICMPv6 messages; Certificate Path Solicitation (CPS) and Certificate Path Advertisement (CPA) as specified in RFC 3971 [10]. Although, SeND was able to prevent malicious attacks on IPv6 neighbor discovery. However, researches have proven [11, 12] that SeND has a drawback like high computation to generate the options especially the CGA option and RSA signature. Thus, it consumes higher computation time. Based on the previous research, SeND mechanism adds significant processing time and it takes 367.59 milliseconds to perform the message verification operation [12]. Hence, if SeND is implemented, its processes i.e. authorization and certificate validation function can add delay and increase complexity during DAD process as highlighted by the researchers [7]. Thus, any malicious host can exploit this mechanism and can cause DoS attack against the SeND mechanism itself by engaging the victim host in message verification processing.

B. *Simple Secure Addressing Scheme (SSAS)*

In order to address the issues with SeND mechanism, another mechanism known as Simple Secure Addressing Scheme (SSAS) was proposed which is considered as an improved version of SeND mechanism on securing ND messages during DAD process in IPv6 network [11]. SSAS introduces alternative addressing scheme by employing elliptic curve cryptography (ECC) algorithm rather than RSA as used by SeND mechanism for address configuration process. In other words, SSAS mechanism is lightweight version of SeND mechanism. In order to protect ND message from spoofing attacks SSAS uses Signature and Timestamp options which are appended to ND messages during DAD process. Although, SSAS has reduced some complexity and resulted in decreased message processing time compared to SeND mechanism. Since this method relies on signature and key exchange processes, hence the complexity issue still exists [12]. Based on the research conducted by Praptodiyono et al. in 2015 [12], SSAS mechanism takes 223.1 milliseconds to generate an interface identifier which is a considerable amount of processing time. Thus, due to its complexity issue, SSAS mechanism can also induce DoS attack on DAD process by delaying the message verification process during address configuration in IPv6 link local network.

C. Trust-ND

Recently, researchers have claimed a lightweight mechanism for DAD process in IPv6 network known as Trust-ND [12]. The main focus of this mechanism has been the complexity of the ND message processing. Trust-ND has significantly reduced the processing time of ND messages during DAD process compared to existing mechanisms such as: SeND and SSAS in IPv6 network. In Trust-ND, message authentication is a result of SHA-1 operation as a message integrity check. Thus, Trust-ND mechanism relies on SHA-1 hash function to satisfy the security requirements. Although, the authors claims that Trust-ND is a lightweight security mechanism for IPv6 DAD process. However, researches [13, 14] have shown that SHA-1 and MD5 hash functions are susceptible to hash collision attacks. Since, Trust-ND's security is based on SHA-1 hash function therefore, any malicious host can exploit this weakness to generate hash collision attack against this mechanism that can cause DoS attack on DAD process in IPv6 network. Thus, due to this security vulnerability Trust-ND might not be a suitable mechanism for IPv6 DAD process.

Due to the constraints possessed by existing security mechanisms as aforementioned. The implementation of the security mechanisms for IPv6 DAD process has been limited. As a result, IPv6 DAD process is still unprotected and prone to be exploited by malicious hosts. Therefore, we proposed a new mechanism known as Secure-DAD to secure ND messages during DAD process. Due to its design, Secure-DAD mechanism can protect NS/NA messages from any kind of exploitation such as: spoofing attack, man-in-the-middle attack (MITM), replay attack or hash collision attacks which are responsible for causing DoS attack during DAD process in IPv6 network. The following Section will explain the design and implementation processes of Secure-DAD mechanism.

IV. Design and Implementation of Secure-DAD Mechanism

In case of IPv6 DAD process, authentication is required to protect NS and NA messages from several types of attacks such as: masquerade, content modification, sequence modification and timing modification which eventually leads to DoS attack [16]. Here, DoS attack relates to the absence of the services i.e. to configure unique IP addresses rather than service unavailability due to flooding attacks. In order to authenticate NS and NA messages, research [17] has recommended using the most appropriate hash function which is resistant to hash collision attacks and can also be faster in computation. Researches [17, 18] have proven that Universal Hashing (UMAC) is efficient algorithm and secure than existing hash functions such as: SHA-1 and MD5. Thus, the most suitable and available hash function algorithm has been selected which can satisfy this security requirement. UMAC can provide message integrity to prevent any tempering with NS and NA messages content as the security requirement. Secure-DAD mechanism is built on UMAC hash function algorithm to ensure that the proposed mechanism is reliable and effective enough to secure a DAD process in IPv6 network.

Secure-DAD mechanism introduces a concept of Secure-tag option which will be appended to each ND message i.e. NS and NA messages exchange between the hosts during DAD process in IPv6 network. This Secure-tag comprises of message authentication code (MAC) to distinguish the valid messages from the fake ones. After the addition of the Secure-tag, these ND messages i.e. NS and NA messages are named as Secured NS and Secured NA messages. Figure 3 and Figure 4 presents the Secured NS and Secured NA messages format respectively.

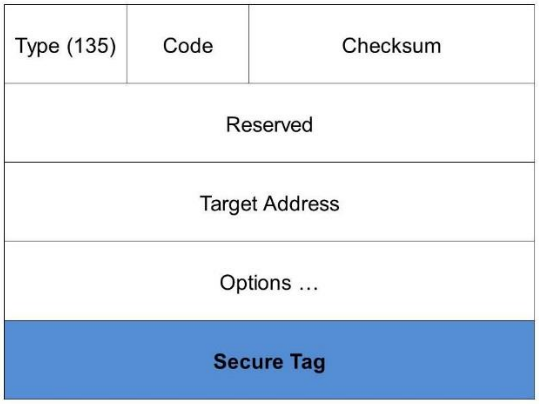


Fig. 3. Secured NS message format

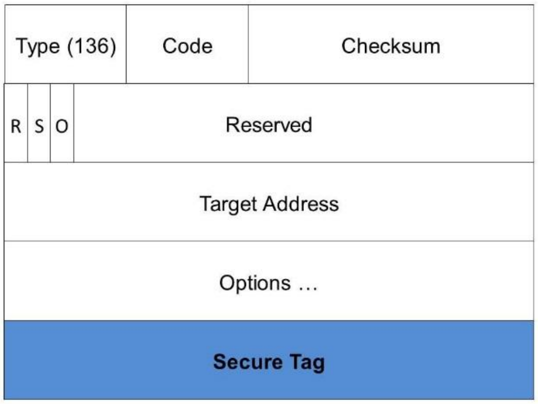


Fig. 4. Secured NA message format

In Secure-DAD mechanism, when a new host performs DAD process it will generate a Secure-tag, appends onto NS message and sends it to multicast address group i.e. FF02::1. Upon receiving NS message existing host(s) will match this Secure-tag option with its self-generated Secure-tag. After the computation process, if these Secure-tags match, then it will perform DAD process and can reply via Secured NA message i.e. NA message appended with Secure-tag. Similarly, upon receiving the NA message, new host performs the same procedure i.e. matching of Secure-tags, else if no match of Secure-tags is found then new host will simply discard the received NA message. Hence, in this manner, new host can perform DAD process successfully. Thus, new host can configure a unique IPv6 link local address. Figure 5 illustrates the Secure-tag generation and verification processes between the hosts in IPv6 link local network.

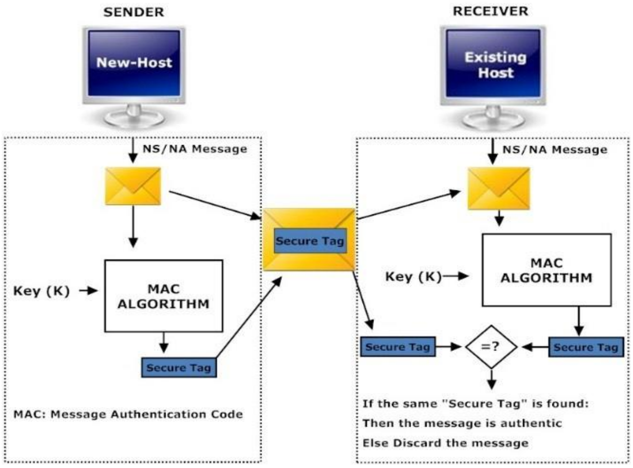


Fig. 5. Secure-tag generation and verification process

V. Test-Bed Setup Environment

In order to evaluate the performance of secure-DAD mechanism in terms of processing time and effectiveness a Test-bed setup has been deployed at NAv6 research Centre in University Science Malaysia (USM). Figure 6 shows the topology of the Test-bed setup environment.

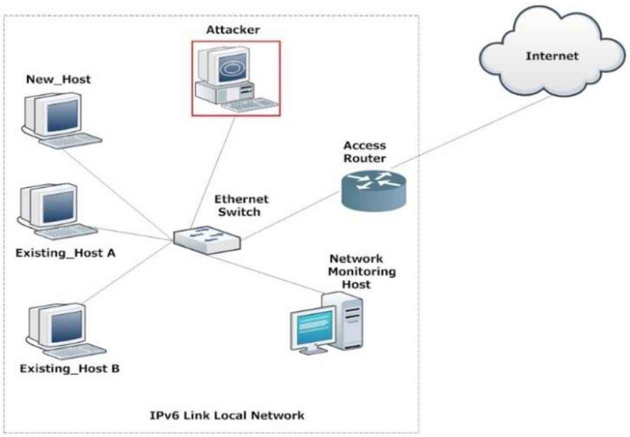


Fig. 6. Test-bed setup environment

The attack could be coming from any type of host's i.e. Windows, Linux etc., since we are using Kali [19] for that purpose attacker host is Linux host. A packet capturing tool known as Wireshark [20] has been used to capture and analyse the network traffic. Moreover, the hardware and software specifications have been selected based on the availability and support for IPv6 environment at NAv6 research Centre to conduct the experiments successfully. The details of the required hardware and software specifications for Test-bed environment setup are presented in Table 1 and Table 2 respectively.

TABLE I. Hardware Requirements for the Experiments

Hardware		Details
Computer Hardware @ per (Host)	CPU	Intel® Core™2Duo CPU E6750 @ 2.66GHZ
	Memory	1 GB Ram
	Network Interface Card	Intel® 82579LM Gigabit1 Ethernet LAN 10/100/1000
	Network Patch cables	Digitus UTP Cat5e
Other Network Devices	Switch	Cisco Catalyst 2960 Fast Ethernet
	Access Router	Cisco Router C7200

TABLE II. Software Requirements for the Experiments

Operating System		Role	Tools
Microsoft Windows	Windows 7 Ultimate 64-bit (version: 6.1.7601.17514)	Network Monitoring Host	Wireshark
		New_Host	-
		Existing_Host A	-
		Existing_Host B	-
Linux Distributions	Kali Linux (version 3.18.0-amd64)	Attacker Host	THC IPv6 Attack Toolkit 2.7

VI. Evaluation of Secure-DAD Mechanism

In order to evaluate the proposed Secure-DAD mechanism, Network security experts have specified a standard criterion known as Information Technology Security Evaluation Criteria (ITSEC) [21]. Therefore, ITSEC has been used to assess the Secure-DAD mechanism. ITSEC presented three metrics for evaluation i.e. Operation, Effectiveness and Functionality [22]. According to ITSEC, any security mechanism that can fulfill these three parameters is considered applicable. Since, Secure-DAD is defined to prevent ND messages from any exploitation which can induce DoS attacks on DAD process by Secure-tag option. Hence, the performance of the Secure-DAD mechanism was evaluated based on these recommended criteria as described in the following Section.

VII. Experimental Results and Discussion

Secure-DAD mechanism is implemented based on the Test-bed environment as presented in Figure 6. In order to make sure that the proposed Secure-DAD mechanism works properly and satisfies the security requirements, the implementation was done in two scenarios. The reason behind that was to measure the performance of Secure-DAD in terms of processing time in first scenario and also, the effectiveness, and functionality of the mechanism in second scenario.

A. *Experiments in First Scenario*

In first scenario experiments were conducted to examine the performance of Secure-DAD mechanism in terms of processing time. In order to fulfill these requirements, Secure-DAD was performed on Test-bed environment setup to measure the Secured ND messages processing time i.e.

Secured NS and Secured NA messages between the sender and receiver hosts. In addition, same experiments were also conducted for the standard DAD process and Trust-ND mechanism on same Test-bed environment. The purpose of conducting these experiments on standard DAD, Secure-DAD, and Trust-ND were to obtain the results of NS and NA messages processing time between IPv6 hosts during DAD process in IPv6 link local network. These results were then analyzed by comparing the three mechanisms to justify the performance of Secure-DAD mechanism. The obtained results are discussed in the following sub-section.

B. Results Analysis and Discussion

This section provides the results analysis and discussion of the operation of Secure-DAD compared against the standard DAD and Trust-ND mechanism. The metric to measure the performance of the Secure-DAD operation along with standard DAD process and Trust-ND mechanism is the processing time of received NS and NA messages at the sender (New_Host) and receiver (Existing_Host) during DAD process respectively. The measurement of Secure-DAD processing time was done by subtracting the end time with the start time of the NS and NA messages verification process at the receiving host. Similarly, the same operation was performed with standard DAD and Trust-ND respectively. It was conducted for each of the NS/NA message for 10 (times) experiment. The comparative results are presented in a graphical form as shown in Figure 7 and Figure 8 for standard DAD process, Secure-DAD and Trust-ND mechanism respectively.

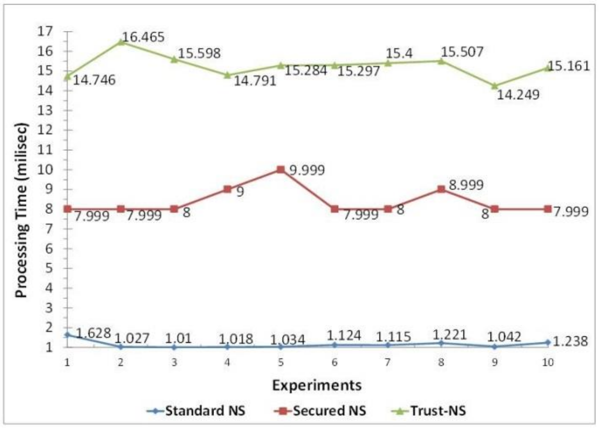


Fig. 7. Comparative NS messages processing time

Figure 7 presents the NS messages processing time at the receiver side i.e. Existing_Host. For each message i.e. Standard NS, Secured NS, and Trust-NS messages experiments were repeated 10 times separately. The purpose for doing this was to find the level of consistency i.e. the average processing time of the NS messages processing time performed for each attempt. Figure 7 also depicts the amount of processing time taken by the three different messages types for each experiment. It shows the level of consistency of the message processing time taken by these message types at the receiver host i.e. Existing_Host.

Table 3 presents the average processing time of the 10 experiments conducted on each message type, as well as the overhead introduced in each Secured NS and Trust-NS messages respectively. The overhead was calculated by putting the standard NS messages average processing time as the baseline. Later, it was compared with Secured NS and Trust-NS messages processing time at the receiver host respectively. Secured NS message processing time is 7.253 milliseconds in average. However, it was also noticed that the Trust-NS message processing time is higher that reaches to 15.250 milliseconds in average. Thus, from the experimental results, it is clear that Secured NS messages consumes less processing time than the Trust-NS messages, which consumes more processing time at the receiver host.

TABLE III. NS MESSAGES PROCESSING TIME AT RECEIVER HOST

Processing Time of NS messages (milliseconds)			
Receiver (Existing_Host)	**Standard NS**	**Secured NS**	**Trust-NS**
Mean	1.146	8.399	15.250
Overhead	Baseline	7.253	14.104

Likewise, sender host i.e. New_Host performs the message verification for all incoming NA messages. The incoming NA message is the response to its NS message sent earlier to Existing_Hosts on a same link to complete the DAD process in IPv6 link local network. Similarly, the sender host i.e. New_Host goes through the same message verification process as performed by the Existing_Host. Therefore, in case of Secured NA message, New_Host verifies the Secure-tag option and its message content. Whereas, in case the incoming message is Trust-NA, it verifies the Trust option and its message content. For standard NA, message processing takes place without the verification of message content. Since standard NA message does not contain any such option to be processed.

Figure 8 depicts the NA messages processing time at the sender side i.e. New_Host. Again for each message type i.e. Standard NA, Secured NA, and Trust-NA messages, individual experiments were conducted 10 times for each mechanism. Figure 8 demonstrates the different processing time for each message types which were carried out 10 times for each experiment. It also presents the level of consistency performed by each message types during the message processing at the sender host i.e. New_Host.

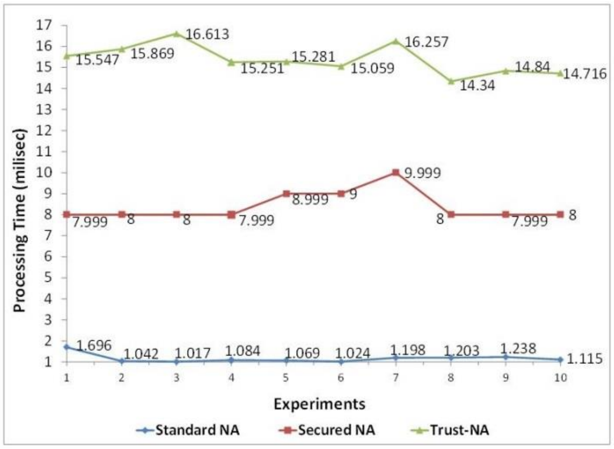


Fig. 8. Comparative NA messages processing time

Table 4 depicts the average processing time consumed by each message types at the sender host i.e. New_Host. Experiments were carried out 10 times on each message types. In addition, to the overhead introduced in average by each Secured NA and Trust-NA messages are also presented. The overhead was estimated by placing the standard NS messages average processing time as a baseline. In this manner, Secured NA and Trust-NA messages processing time were calculated accordingly.

TABLE IV. NA MESSAGES PROCESSING TIME AT RECEIVER HOST

Processing Time of NA messages (milliseconds)			
Sender (New_Host)	**Standard NA**	**Secured NA**	**Trust-NA**
Mean	1.169	8.499	15.377
Overhead	Baseline	7.330	14.208

Table 5 shows the overall processing time differences between the standard DAD, Secure-DAD, and Trust-ND mechanisms. The processing time of ND messages i.e. NS and NA messages between the IPv6 hosts represents the computational efficiency of security mechanism. Therefore, by comparing the processing time of Secure-DAD and Trust-ND mechanisms with the standard DAD as a baseline, effects of these two mechanisms on DAD process in IPv6 network can be distinguished.

TABLE V. OVERALL PROCESSING TIME AT SENDER AND RECEIVER HOSTS

DAD Process	Processing Time (milliseconds)		
	Standard DAD	**Secure-DAD**	**Trust-ND**
Sender (New_Host) NS	1.146	8.399	15.250
Receiver (Existing_Host) NA	1.169	8.499	15.377
Total	2.315	16.898	30.627
Overhead	Baseline	14.583	28.312

The overall processing time of standard DAD, Secure-DAD, and Trust-ND mechanisms are 2.315, 16.898, and 30.627 milliseconds respectively. Hence, the total overhead introduced by Secure-DAD mechanism is 14.583 milliseconds in average. Whereas, Trust-ND mechanism is 28.312 milliseconds in average. Thus, the overhead introduced by Secure-DAD is lesser as compared to Trust-ND mechanism.

Table 6 depicts the saved processing time on the implementation of Secure-DAD against Trust-ND mechanism. Secure-DAD is able to save time up to 13.729 times, which means processing time reduction of 45.1% compared to Trust-ND correspondingly for NS and NA messages processing time during address verification process between hosts in IPv6 link local network.

TABLE VI. PROCESSING TIME SAVED BY SECURE-DAD

DAD Process	Processing Time (milliseconds)		Saving Time (milliseconds)
	Trust-ND	**Secure-DAD**	
Sender (New_Host) NS	15.250	8.399	6.851
Receiver (Existing_Host) NA	15.377	8.499	6.878
Total	30.627	16.898	13.729

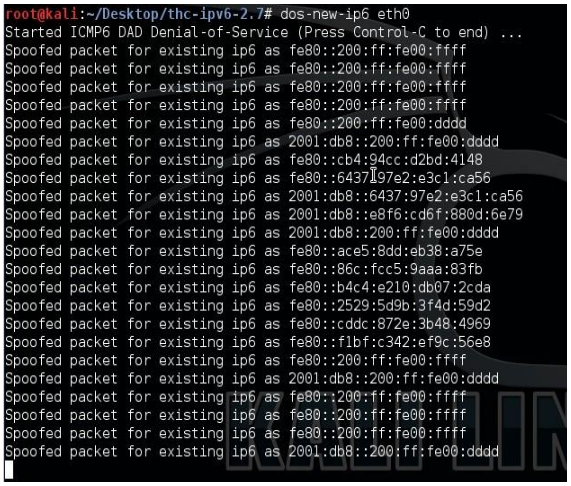
Thus, from the results it is clear that the proposed Secure-DAD mechanism is able to reduce the level of complexity i.e. the processing time of NS and NA messages verification at the hosts during DAD process in IPv6 link local network. This is in contrast to the Trust-ND mechanism and other existing mechanism such as: SeND, SSAS that possess the high level of complexity as stated by the researchers [12].

C. *Experiments in Second Scenario*

The second scenario was conducted to validate the effectiveness of Secure-DAD under the attacking situation. This scenario was examined to ensure that Secure-DAD mechanism is capable of protecting NS and NA messages from fabricating during DAD process which can eventually causes DoS attack. In order to test the Secure-DAD, the attacking approach was performed by running dos-new-ip6 attack tool [23]. The purpose of carry out denial of service attack was to measure the effectiveness of Secure-DAD mechanism to satisfy the functionality requirement under attack condition this was done by using the dos-new-ip6 attack tool.

- Attack Generation on DAD Process

The main purpose of attacking a New_Host is to cause the host initialization failure. In order to achieve this aim, Attacker host uses dos-new-ip6 tool to generate a NA message to answer whatever tentative address is being generated by the New_Host. This is intended to cause DAD process failure which can deny New_Host to obtain a unique IPv6 address. Figure 9 depicts the DoS-on-DAD attack generation against DAD process in IPv6 network.

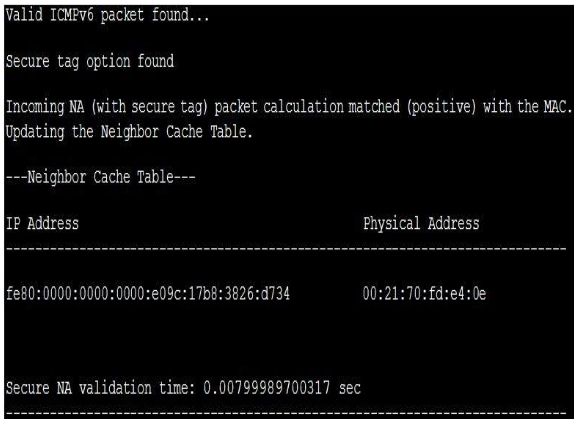


```
root@kali:~/Desktop/thc-ipv6-2.7# dos-new-ip6 eth0
Started ICMP6 DAD Denial-of-Service (Press Control-C to end) ...
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as fe80::200:ff:fe00:dddd
Spoofed packet for existing ip6 as 2001:db8::200:ff:fe00:dddd
Spoofed packet for existing ip6 as fe80::cb4:94cc:d2bd:4148
Spoofed packet for existing ip6 as fe80::6437:97e2:e3c1:ca56
Spoofed packet for existing ip6 as 2001:db8::6437:97e2:e3c1:ca56
Spoofed packet for existing ip6 as 2001:db8::e8f6:cd6f:880d:6e79
Spoofed packet for existing ip6 as 2001:db8::200:ff:fe00:dddd
Spoofed packet for existing ip6 as fe80::ace5:8dd:eb38:a75e
Spoofed packet for existing ip6 as fe80::86c:fcc5:9aaa:83fb
Spoofed packet for existing ip6 as fe80::b4c4:e210:db07:2cda
Spoofed packet for existing ip6 as fe80::2529:5d9b:3f4d:59d2
Spoofed packet for existing ip6 as fe80::cddc:872e:3b48:4969
Spoofed packet for existing ip6 as fe80::f1bf:c342:ef9c:56e8
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as 2001:db8::200:ff:fe00:dddd
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as fe80::200:ff:fe00:ffff
Spoofed packet for existing ip6 as 2001:db8::200:ff:fe00:dddd
```

Fig. 9. Carrying out DoS-on-DAD attack

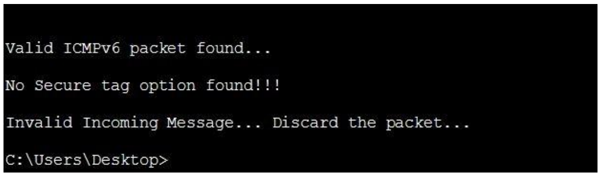
- Prevention Approach

In order to prevent the occurrence of DoS-on-DAD attack, New_Host was enabled with Secure-DAD mechanism that wants to join the IPv6 link local network. To prevent itself from DoS attack, it performed the validation check on every incoming NA message as aforementioned in Section 4. New_Host discarded all ND message such as; in coming NA messages except Secured NA messages appended with Secure-tag option from the sender (Existing_Host), while conducted Secure-tag matching process for all incoming NA messages with its self-generated Secure-tag. For instance, when New_Host received any NS message from the Existing_Hosts, It performed Secure-tags matching process. It entertained only those incoming NA messages that contains Secure-tag option while rest of the incoming NA messages were discarded. Figure 10 and Figure 11 presents the Secure-tag validation performed by New_Host upon receiving the valid Secured NA message from the valid host and fake NA message from an attacker host respectively.



```
Valid ICMPv6 packet found...

Secure tag option found

Incoming NA (with secure tag) packet calculation matched (positive) with the MAC.
Updating the Neighbor Cache Table.

---Neighbor Cache Table---

IP Address                                  Physical Address
----------------------------------------------------------------------------

fe80:0000:0000:0000:e09c:17b8:3826:d734         00:21:70:fd:e4:0e



Secure NA validation time: 0.00799989700317 sec
----------------------------------------------------------------------------
```

Fig. 10. Secure-tag validation for incoming NA message



```
Valid ICMPv6 packet found...

No Secure tag option found!!!

Invalid Incoming Message... Discard the packet...

C:\Users\Desktop>
```

Fig. 11. Secure-tag validation process failure

Hence, from the experimental tests and results, it is clear that the Secure-DAD is an improved mechanism both in terms of processing time and effectiveness to prevent DoS attacks during DAD process in IPv6 link local network. The results have also proven that the Secure-DAD consumes less processing time to perform DAD process as compared with the existing mechanisms such as; SeND, SSAS, and Trust-ND. Moreover, Secure-DAD is effective enough to prevent DoS attack on DAD process. Figure 12 depicts the comparatives analysis of all mechanisms (SeND, SSAS, Trust-ND, and Secure-DAD) in terms of processing time to perform DAD process in IPv6 link local network. Thus, Secure-DAD is a suitable mechanism for IPv6 hosts to perform a secure link local communication in IPv6 network.

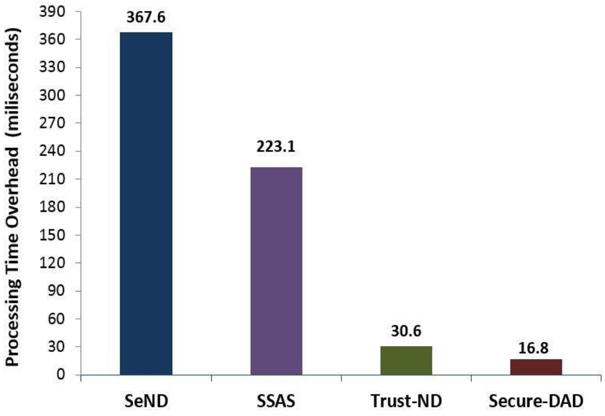


Fig. 12. Comparative results of Secure-DAD with existing mechanisms in terms of processing time overhead (SeND and SSAS processing time results were adopted from [12])

VIII. Conclusion and Future Work

This paper presented an improved mechanism to prevent DoS attack on DAD process in IPv6 network. A Test-bed was designed to allow the authors to evaluate the effectiveness of the mechanism by carrying out DoS attacks and comparing the performance of Trust-ND and Secure-DAD mechanisms. The experimentations were conducted on standard DAD, Secure-DAD, and Trust-ND mechanisms to justify the performance of Secure-DAD. The results showed that Secure-DAD consumed less processing time compared to Trust-ND mechanism. Moreover, Secure-DAD possessed less complexity compared to existing mechanisms such as; SeND, SSAS, and Trust-ND. Therefore, Secure-DAD is computationally efficient compared to existing mechanisms. In addition, experimented results also proved that the Secure-DAD mechanism is resistant to different types of attacks which can induce DoS attacks directly or indirectly on DAD process in IPv6 link local network i.e. effective and functional.

Hence, from the experimental tests and results, it was evaluated that the Secure-DAD mechanism not only performed better in terms of processing time, but also was effective and functional during attack conditions. Currently, the Secure-DAD mechanism was implemented on a small scale private IPv6 network. Therefore, our future work will be to optimize the Secure-DAD mechanism so that it can be applicable for the large scale public area IPv6 network.

ACKNOWLEDGMENT

This research was supported by National Advanced IPv6 Centre (NAV6), Universiti Sains Malaysia (USM) and Ministry of Higher Education Malaysia.

REFERENCES

[1] Thomson S, Narten T, Jinmei T. IPv6 Stateless Address Auto-configuration. Internet RFC 4862, 2007.

[2] Deering S, Hinden R. Internet protocol version 6 (IPv6) specification. Internet RFC 2460, 1998.

[3] Li, S., Da Xu, L., & Zhao, S. The internet of things: a survey. Information Systems Frontiers, Springer, Science & Business Media, vol. 17(2), pp. 243-259, 2015.

[4] Botta, A., de Donato, W., Persico, V., & Pescapé, A. Integration of cloud computing and internet of things: a survey. Future Generation Computer Systems, Elsevier, 56, pp.684-700, 2016.

[5] Rehman SU, Manickam S. Significance of duplicate address detection mechanism in Ipv6 and its security issues: A survey. Indian Journal of Science and Technology, vol. (8)30, 2015.

[6] Narten T, Simpson, WA, Nordmark E, Soliman H., Neighbor discovery for IP version 6 (IPv6), 2007.

[7] AlSa'deh A, Meinel C. Secure neighbor discovery: Review, challenges, perspectives, and recommendations. IEEE Security & Privacy, vol. 10, pp. 26-34, 2012.

[8] Conta A, Gupta M. Internet control message protocol (ICMPv6) specification. Internet RFC 4443, 2006.

[9] Dawood, H. IPv6 Security Vulnerabilities. International Journal of Information Security Science, vol. 1(4), pp.100-105, 2012.

[10] Arkko J, Kemp f J, Zill B, Nikander P. Secure neighbor discovery (SEND). Internet RFC 3971, 2005.

[11] Rafiee H, Meinel C. SSAS: A simple secure addressing scheme for IPv6 autoconfiguration. Eleventh Annual IEEE International Conference on Privacy, Security and Trust (PST), pp. 275-282, 2013.

[12] Praptodiyono S, Murugesan R K, Hasbullah IH., Wey CY, Kadhum MM, Osman A. Security mechanism for IPv6 stateless address autoconfiguration. 2015 IEEE International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), pp. 31-36, 2015.

[13] Andreeva E, Mennink B, Preneel B. Open problems in hash function security. Designs, Codes and Cryptography, vol. 77, pp. 611-631, 2015.

[14] Bhargavan K, Leurent G. Transcript collision attacks: Breaking authentication in TLS, IKE, and SSH. NDSS, 2016.

[15] Rehman SU, Manickam S. Denial of Service Attack in IPv6 Duplicate Address Detection Process. International Journal of Advanced Computer Science & Applications, vol. 7, pp. 232-238, 2016.

[16] Moore D, Shannon C, Brown D J, Voelker GM, Savage S. Inferring Internet denial-of-service activity. ACM Transactions on Computer Systems (TOCS), vol. 24. Pp. 115-139, 2006.

[17] Shoup V, fast and provably secure message authentication based on universal hashing. In Advances in Cryptology—CRYPTO'96, pp. 313-328, 1996.

[18] Krovetz T. UMAC: Message authentication code using universal hashing. Internet RFC 4418, 2006.

[19] Kali Linux Penetration Testing and Ethical Hacking Linux Distribution. https://www.Kali.org.

[20] V. Ndatinya, Z. Xiao, V. R. Manepalli, K. Meng, and Y. Xiao, "Network forensics analysis using Wireshark," International Journal of Security and Networks, vol. 10(2), pp. 91–106, 2015.

[21] Woodcock, J., Stepney, S., Cooper, D., Clark, J., & Jacob, J., The certification of the Mondex electronic purse to ITSEC Level E6. Formal Aspects of Computing, vol. 20(1), pp. 5-19, 2008.

[22] Saleem S, Popov O, Dahman R. Evaluation of security methods for ensuring the integrity of digital evidence. 2011 IEEE International conference on Innovations in information technology (IIT), pp. 220-225, 2011.

[23] THC-IPv6 Attack Tool-kit. https://www. aldeid. Com/wiki/THC-IPv6-Attack-Toolkit.

Web Server Performance Evaluation in a Virtualisation Environment

Performance evaluation of Web Server

Manjur Kolhar
Dept. Computer Science
Prince Sattam Bin Abdulaziz University,
Wadi Ad Dawaser, KSA

***Abstract*—Operational and investment costs are reduced by resource sharing in virtual machine (VM) environments, which also results in an overhead for hosted services. VM machine performance is important because of resource contention. If an application takes a long time to execute because of its CPU or network, it is considered to be a failure because if many VMs are running over a single hardware platform, there will be competition for shared resources, e.g., the CPU, network bandwidth, and memory. Therefore, this study focuses on measuring the performance of a web server under a virtual environment and comparing those results with that from a dedicated machine. We found that the difference between the two sets of results is largely negligible. However, in some areas, one approach performed better than the other.**

Keywords—Cloud computing; virtual machine; resource sharing; latency sensitive; web server; multi-tier application

I. Introduction

Cloud computing allows the user to store and access data over the Internet using virtualisation technology. Virtualisation is software that serves as an intermediary between the physical network and the cloud. Furthermore, virtualisation allows a cloud service provider (CSP) to run multiple operating systems using a VM over a single hardware system, which reduces operating and investment costs. Virtualisation and cloud computing differ in that virtualisation runs on hardware, whereas cloud computing is a service resulting from the virtualisation [1–2]. CSPs provide these services through a co-tenant scheme. Services or applications running in a virtualised environment demand more processing power from the host hardware system [3-4]. Additionally, virtualisation overhead may occur because of the processing time for various services and tenant schemes. Hence, it is necessary to measure the behavior of an application under various virtual environmental conditions before moving an application permanently to the cloud. Latency-sensitive applications suffer because of resource, network, and CPU sharing, eliminating these virtualisation benefits [5].

In this study, we host a web server on a VM to measure latency-sensitive elements influencing its performance. The performance of a client-side application, i.e., a web browser, is also measured because it is involved in request and response transactions that involves the sharing of major resources, including memory, network I/O virtualisation, and CPU.

In this study, we evaluate a web server's performance in a virtualisation environment and compare it with that of a dedicated web server running on a machine without a VM. We compare with a baseline system for application performance and security resource consumption – that is, a web server and a web server on the client machine – because performance and resource consumption depend on the virtualisation configurations. Additionally, to secure hosted applications, CSPs install security patches on the cloud. This setup induces more latency in the application environment. Therefore, we present the results of our experiments to answer following questions:

- How does the virtualised web server's performance compare to the performance of a dedicated server, including request and response time?
- How is web browser performance affected when multiple tiers are used?
- As the number of multi-tiered applications increases, how is the web browser served to users under the influence of VM (e.g., during content loading)?

This manuscript is organised as follows. First, in Section 2, we summarise past literature. In Section 3, we discuss the methodology for measuring the CPU, network, and other modules that influence the performance of the virtual and dedicated environments. In Section 4, we evaluate the given methodology, and finally, in Section 5, we outline our conclusions.

II. Literature Survey

Real-time data transmission over a network is built on the assumption that the response and request primitives are executed in a specific time, no messages are lost, multiple running applications do not interfere with each other, and transactions are not influenced. However, these transactions are a source of unpredictable patterns of communication over a network. Moreover, these sets of network communication patterns are not communicated in the same fashion on VM operating systems owing to the multi-tenancy concept of cloud computing and a higher consolidation of resource sharing. Furthermore, sharing resources such as CPU, memory, and network adapters between VM tenants and applications makes it more difficult to provide steady service and predictable network performance. However, VMs are capable of executing concurrently and with the support of underlying hardware and hypervisor.

Running web-based services or a website on these VMs is also a major challenge because they constantly read, write, and update data. These transactions are time-bounded requests and reply primitives and may not be served properly under latency-sensitive environments induced by the resource sharing in cloud computing; it is possible for individual transactions to overlook their own latency requirements [6].

Recently, cloud computing has been a primary focus for numerous computing applications. Using virtualisation, substantial growth has been achieved for many different workloads in both web-based services and cloud clients. Since the establishment of cloud computing for hosting services, researchers have been evaluating cloud performance under virtualisation. Very recently, the usefulness of general-purpose graphical programming units (GPUs) was measured; it was found that the GPU can greatly influence the performance of hosted services by means of a peripheral component interface [7].

The authors of [8] have evaluated network virtualisation overheads in the Xen environment using different workloads and under different configurations. A micro-level web server stressed the overall networking system. The stress test involved data transmission and connection establishment and closing [8]. The Xen VM was used for monitoring CPU usage of different VM overheads in the device driver domain due to I/O of VM, which was intended to quantify and measure the overhead caused due to I/O-intense jobs [8].

VM clusters based on I/O communication have improved and optimised network usage in data centres [10]. This study used a greedy algorithm to guarantee that the migration of lower-priority placement decisions was swift, thus making it suitable for large data centres. To maintain a service level agreement (SLA), an algorithm is proposed that is based on adaptive utilisation thresholds [11]. To reduce memory footprints, page-sharing models were introduced for VM co-hosting [12].

An online self-reconfiguration-based reallocating framework for VMs is proposed in [13]. The framework accurately forecasts the workloads of VM requests with Brown’s quadratic exponential smoothing. Linear programming and heuristics are used for VM migration, which helps in prioritising VMs with fixed capacity [14]. In [15], an energy-aware heuristic framework is proposed for VMs to maintain SLAs and to use minimum power for maximum utilisation.

In [16], a VM resource demand predictor is proposed for allocating cloud applications. Researchers proposed a heuristic scheduling VM with adaptive resource allocation for reducing the number of physical machines.

Researchers also performed live migration of multiple VMs to reduce the traffic load on network links. Migration is carried out using distributed reduplication of VMs’ memory images [17]. In [18], authors studied virtual switching overhead on a server and proposed virtual switching-aware algorithms.

In [19], a novel analytical model is proposed that is built on a queuing network to measure the performance of virtualised multi-tier applications. The effectiveness of the proposed model is assessed by a series of comprehensive trials of different configurations of multi-tier applications.

However, none of the above literature considers security features, the migration of server security along with the web server, or of application to the VM. Hence, our work is focused on these issues. A single powerful hardware system may host multiple VMs; these VMs compete for network adapters. Hence, these virtualisations environments induce overhead because of network I/O virtualisation. As with network interface cards and memory, the CPU must be shared among the hosted VMs. Therefore, the CPU also induces latency for the hosted applications.

III. Experimental Setup

Figure 1 shows the virtualisation and dedicated environment of our testbed, hereafter referred to as our testbed. We run experiments in two systems with an identical setup. We compare the performance of virtual box with that of a dedicated system. For a virtualised environment configuration, our physical machine may host one or more virtual boxes because we are interested in multi-tiered applications. Similarly, the dedicated server is also hosting multitier applications.

Default server installations on the testbed have default OS configurations, system services, and network services that are not secure. Unnecessary services and their ports are open and not used for the testbed environment. Hence, these services and ports are closed to avoid malicious intrusion. Our testbed environment needs remote access; it has secure remote access using tunnelled and encrypted protocols. We have enabled and allowed file and network service permissions and privileges on our testbed. To secure our web-based application, we have updated security patches to the latest versions.

Web-based applications are being developed for various scenarios ranging from small- to large-scale business environments. We are running time-sensitive web applications developed with the help of the Apache web-server, Java, and MySQL databases. A major threat to the availability of web applications comes from distributed denial of service attacks, which is the overloading of fake requests to a web server so as to deny a legitimate user access. Such threats work at both the lower (TCP/IP) and upper (application) layers of a network. To run a web server on the Internet, the network administrator should be well-versed in these threats. Hence, our testbed is also armed with security solutions to avoid such a threat.

To begin our experiments a dedicated testbed server was used. On this server, we have utilised a time-sensitive web application that uses a Tomcat server and the Java programming language. Our web application is a voice-over-Internet protocol (VoIP), which is a real-time media transmission protocol. This server requires end-users to register before using the web server for making audio and video calls over the Internet. We measured response times by registering more than 100 users at a time on the web server. During registration, the web server must perform a number of tasks;

first, it takes the user name and password from the user, cross-verifies to authenticate them, and replies to the user with a "200" message code. Once the user sends the registration requests to the server, the client request crosses a network connection from the client to the server. We implemented the above experimental testbed in our university lab; thus, decreasing the time required for a packet to travel from source to destination. Furthermore, we are very much concerned with the response messages originating from the server and not on the network path.

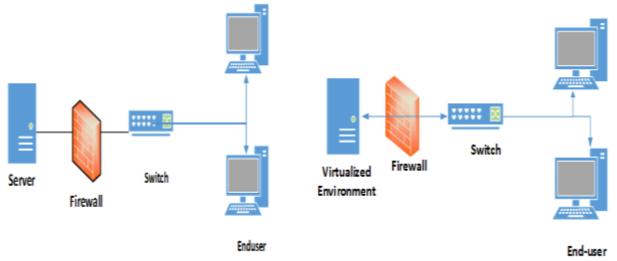


Fig. 1. Dedicated server (Left) Virtualized server environment (Right)

IV. Performance Study

We performed testbed server experiments that exercise our network and database traffic in order to estimate the CPU, network, and memory usage caused by the registration process of a web server application. All experiments were performed on an OptiPlex 3020 Micro PC. For these measurements, we used Linux 2.6.8.1, as mentioned in Table 1.

TABLE I. Configuration for the Experimental Setup

Node	Hardware configuration	
	Type	*CPU*
PC1@ university campus	OptiPlex 3020 Micro	Intel ® Pentium G3250T Processor (Dual Core, 3 MB, 2.8 GHz w/HD Graphics)
PC2@ university campus	OptiPlex 3020 Micro	Intel ® Pentium G3250T Processor (Dual Core, 3 MB, 2.8 GHz w/HD Graphics)

We began the first group of web server performance evaluations, measuring the number of requests served by the web server, its response time in milliseconds, and its throughput measured in bytes per seconds. The measurements were performed under a variable number of registration requests from clients. Web server performance under high workloads is network-bounded and under low workloads, CPU-bounded. Hence, we measured both conditions to evaluate the CPU and network interfaces.

To evaluate the CPU overhead of a dedicated server of varying web traffic, we used an Apache Tomcat HTTP server running on the testbed and a PC for sending VoIP registration requests. We used the session initiation protocol (SIP) tester to generate VoIP traffic registration requests. This tool issues a variable number of registration requests and is specifically designed for evaluating VoIP servers. We can increase the registration request rate until we receive a low reply rate from the server; that is, until the server becomes saturated. We formed a group of SIP server workloads, each generating registration requests from a variable number of clients: 10 to 100 clients per second, in steps of 10 clients.

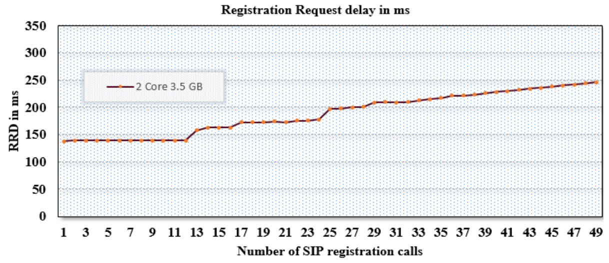


Fig. 2. Dedicated server performance

The maximum load applied to the web server is 100 registrations per second. Figure 2 shows that this is equivalent to a CPU load of 50 requests per second. Similarly, minimum throughput was achieved under a workload with 2 requests per second, a value related to the applied load. Figure 2 shows the performance of the web server on a dedicated machine. Figure 3 shows the performance of the web server on a VM.

According to the International Telecommunication Union (ITU), end-to-end, one-way delay in media transmission is 400 ms. However, there exist different delays for different codec algorithms. Media transmission protocols should abide by this law to successfully provide VoIP service. However, our testbed performance in both experiments found the end-to-end, one-way delay in media transmission to be much less than 400 ms, as shown in Figures 2 and 3.

Hence, we conclude that registration of VoIP under VMs is considered acceptable. Furthermore, we have also measured the CPU load, requests (using the GET method) made on the web server by a virtual user (VU), and bandwidth between the web server and client. We found that the TCP connections made per second are proportional to the number of VUs. We noticed that both bandwidth and CPU are directly proportional to the requests made to the web server in an attempt to obtain service from it. Figures 4 and 5 illustrate trace tests on dedicated and VM web servers, respectively. In the testbed, the CPU performance of a VM is higher than that of the dedicated machine, as expected. However, the difference is negligible and depends on bandwidth, because requests and responses are I/O-bound and hence, the CPU is involved in I/O requests. For 8000 requests, only 4% of the CPU is consumed under a VM, whereas this percentage is 3.26% in the case of a dedicated machine. The difference is further reduced with proper usage of para-virtualised devices to services in the VM.

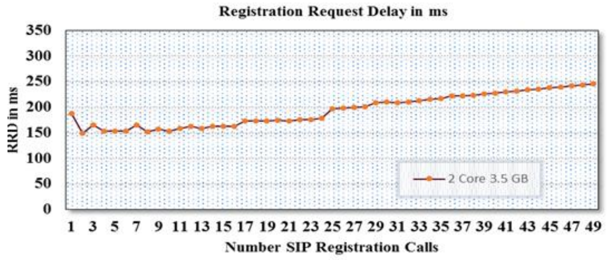


Fig. 3. Virtualized web-server performance

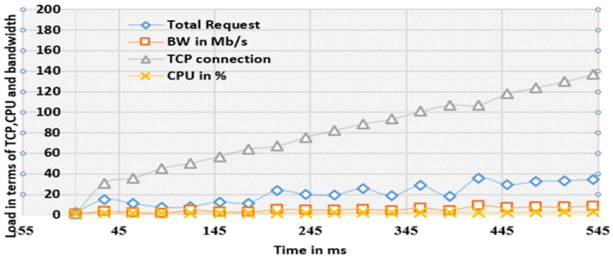


Fig. 4. Dedicated web browser performance

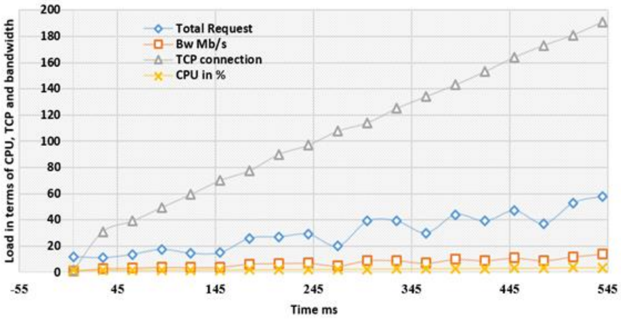


Fig. 5. Dedicated web browser performance

The performance evaluation of our testbed setup has so far been conducted in situations in which there is barely any CPU conflict from multiple VMs. However, in real deployment setups, the shared environment of many VMs in a single host efficiently utilises existing resources. Special physical machine access, enabled by the latency feature, allows the VM to achieve better results because VMs use virtual network interface cards, virtual kernels, and I/O for network-based operations. These physical elements are accessed by the VM through software. If the VM does not obtain the physical machines' power, then VM performance may degrade. In real deployment, multi-tier applications are deployed on different machines. Similarly, it is better to host multi-tier applications on different VMs rather than on a single VM because physically separated applications, such as an application server, database, or other business logic modules, are not directly reachable by hackers at a single machine. Apart from this security benefit, I/O and CPU load are also equally distributed on each of the modules of multi-tier applications. Web browsers load page elements sequentially. These elements include scripts (in HTML, PHP, or other scripting languages), style sheets, and images. However, all these elements are not accessed or downloaded to the web browser at once. Browsers open a limited number of HTTP and TCP connections based on the referenced web page on the server, because of their capacity to load only a limited amount of data per second. Furthermore, the GET and POST methods, respectively, are used to fetch and send data from the server. Therefore, these methods are expensive and are economical models for the CSP. At the same time, these methods are also critical in the performance of any website.

We have measured these methods in our testbed; our website is made from HTML, JavaScript, CSS style sheets, images, and Flash. As our testbed does not have a domain name service (DNS) server, we do not have DNS or an Internet service provider. In table 2, there are two critical methods, namely DNS and Secured Socket Layer (SSL) negotiation. These have consumed much less CPU and bandwidth in both testbed environments. Acquiring a DNS is time consuming in the first instance only. Loading HTML and the corresponding referenced pages are completely network-based operations and incur time costs of more than 2.2 ms and 2.4 ms in dedicated and VM environments, respectively. I/O operations, such as Java scripts, need the CPU and network; therefore, during execution of Java scripts, we noticed bandwidth and CPU consumption of 2.5 ms and 0.5 ms on the VM and dedicated machine, respectively. However, loading images cost more on a dedicated machine than on a VM. Congestion avoidance algorithms are used to control exponential reduction when congestion occurs because of the behaviour of the TCP protocol.

TABLE II. BROWSERS ENVIRONMENT

Server	**Dedicated and Virtualisation**	
	Dedicated	***Virtualisation***
DNS	0.2	0.25
Connect	0.4	0.42
SSL	0.7	0.7
HTML	2.5	2.6
JS	0.75	0.67
CSS	0.35	0.2
image	3.2	2.6
flash	-	--
font	0.2	0.6

V. CONCLUSION

The performance evaluation of web servers and browsers shows that latency-sensitive applications have successfully run without major delay. However, on some occasions, congestion avoidance has caused some issues to both environments because of the built-in features of the TCP protocol. Some latency sensitivity features provided by major VM vendors can be used to improve performance of hosted services on the Internet cloud.

REFERENCES

[1] M. Armbrust et al., "A view of cloud computing," Commun. ACM, vol. 53, no. 4, pp. 50–58, 2010.

[2] T. Ma, Y. Chu, L. Zhao, and O. Ankhbayar, “Resource allocation and scheduling in cloud computing: Policy and algorithm,” *IETE Techn. Rev.*, vol. 2;31, no.1, pp. 4–16, Jan. 2014.

[3] W. D. Mulia, N. Sehgal, S. Sohoni, J. M. Acken, C.L. Stanberry, and D. J. Fritz, “Cloud workload characterization,” *IETE Techn. Rev.*, vol. 1;30, no. 5, pp. 382–397, Sep. 2013.

[4] M. Kolhar, M. Abu-Alhaj, S. M. Abd El-atty, "Cloud Data Auditing Techniques with a Focus on Privacy and Security," IEEE Security & Privacy, vol. 15, no. 1, pp. 42-51, Jan.-Feb. 2017.

[5] M. Kolhar, S. Abd El-atty, Mohammed Rahmath, “Storage allocation scheme for virtual instances of cloud computing”. Neural Computing and Applications, pp.1-8, Dec-Jan 2016. 1-8.

[6] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.

[7] A. J. Younge, et al., "Evaluating GPU passthrough in Xen for high performance cloud computing," in Parallel & Distributed Processing Symposium Workshops (IPDPSW), 2014.

[8] A. Menon et al., "Diagnosing performance overheads in the Xen virtual machine environment," in *Proc. 1st ACM/USENIX Int. Conf. Virtual Execution Environments*, 2005.

[9] L. Cherkasova and R. Gardner. "Measuring CPU overhead for I/O processing in the Xen virtual machine monitor," in *USENIX Annual Tech Conf.,* vol. 50, 2005.

[10] D. Kakadia, N. Kopri, and V. Varma, "Network-aware virtual machine consolidation for large data centers," in *Proc. 3rd Int. Workshop on Network-Aware Data Management,* 2013, pp. 6.

[11] A. Beloglazov and R. Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," in *Proc. 8th Int. Workshop on Middleware for Grids, Clouds and e-Science*, vol. 4, 2010.

[12] M. Sindelar, R. K. Sitaraman, and P. Shenoy, "Sharing-aware algorithms for virtual machine colocation," in *Proc. 23rd Annu. ACM Symp. Parallelism in Algorithms and Architectures*, pp. 367–378, 2011.

[13] H. Mi, H. Wang, G. Yin, Y. Zhou, D. Shi, and L. Yuan, "Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers," in *IEEE Int. Conf. Services Computing*, pp. 514–521, 2010.

[14] T. C. Ferreto, M. A. S. Netto, R. N. Calheiros, and C. A. F. De Rose, "Server consolidation with migration control for virtualized data centers," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1027–1034, 2011.

[15] Z. Cao and S. Dong, "An energy-aware heuristic framework for virtual machine consolidation in cloud computing," *J. Supercomputing*, vol. 69, no. 1, pp. 429-451, 2010.

[16] Q. Huang, S. Su, S. Xu, J. Li, P. Xu, and K. Shuang, "Migration-based elastic consolidation scheduling in cloud data center," in *33rd IEEE Int. Conf. Distributed Computing Systems Workshops*, 2013, pp. 93–97.

[17] U. Deshpande, U. Kulkarni, and K. Gopalan, "Inter-rack live migration of multiple virtual machines," in *Proc. 6th Int. Workshop Virtualization Technologies in Distributed Computing Date*, pp. 19–26, 2012.

[18] M. Li, J. Bi, and Z. Li, "Improving consolidation of virtual machine based on virtual switching overhead estimation," *J. Network and Computer Appl.*, 2015.

[19] K. R. Zadeh, A.L. Morteza, P. Kabiri, and B. Javadi, "Performance modeling and analysis of virtualized multi-tier applications under dynamic workloads," *J. Network and Computer Appl.*, vol. 56, pp. 166–187, 2015.

A Multi-Threaded Symmetric Block Encryption Scheme Implementing PRNG for DES and AES Systems

Adi A. Maaita
Department of Software Engineering
Faculty of Information Technology, Isra University
Amman, Jordan

Hamza A. Alsewadi
Faculty of Information Technology
Middle East University
Amman, Jordan

***Abstract*—Due to the ever-increasing efficiency of computer systems, symmetric cryptosystem are becoming more vulnerable to linear cryptanalysis brute force attacks. For example, DES with its short key (56 bits) is becoming easier to break, while AES has a much longer key size (up to 256 bits), which makes it very difficult to crack using even the most advanced dedicated cryptanalysis computers. However, more complex algorithms, which exhibit better confusion and diffusion characteristics, are always required. Such algorithms must have stronger resistance against differential and linear cryptanalysis attacks. This paper describes the development of an algorithm that implements a pseudo random number generator (PRNG) in order to increase the key generation complexity. Experimental results on both DES and AES cryptosystems complemented with the PRNG have shown an average improvement of up to 36.3% in the avalanche error computation over the original standard systems, which is a considerable improvement in the time complexity of both systems.**

Keywords*—*Computer Security; Symmetric cryptography; DES; AES; pseudo random number generators

I. Introduction

Governments, banks, universities, and regular individuals are sending and receiving colossal amounts of digital data over networks and through other digital means non-stop. The ever flowing torrent of data holds information of varying levels of importance and sensitivity, such of which is determined by the purpose to which it will be put to use by its sender and receiver, and the damage which results from it falling into the wrong hands.

Keeping government, industrial, financial, and personal secrets safe is a paramount concern in a world controlled through digital communications and integrated data storage. Secrets flow from one computer to another until they reach their designated destinations. But what if those secrets were intercepted?

Encryption is an ancient solution designed to protect information which can be intercepted by those who were not meant to receive it. Many algorithms were developed over thousands of years for that purpose. In the digital age, encryption algorithms are classified into symmetric algorithms (secret-key algorithms), and asymmetric algorithms (public-key algorithms) [1, 2]. Symmetric algorithms require both the sender and the receiver of encrypted data to have the same key which will be used for both encryption and decryption, while for asymmetric algorithms, the key used to perform encryption of some data is different from the key which will be used to decrypt that data.

This work is concerned with the enhancement of the secret key generation process using random number generator, that will be used with symmetric cryptographic systems, in particular data encryption standard (DES) and advanced encryption standard (AES). Hence only these two systems will be reviewed together with pseudo random number generators in the following sections. After the brief introduction and the literature review presented in sections 1 and 2, section 3 will include the methodology of the proposed algorithm. Section 4 lists out the obtained results. Section 5 provides a comprehensive discussion of the obtained results. Finally, section 6 concludes the work.

II. Literature Review

This paper is concerned with the improvement of two widely used symmetric cryptosystems: the Data Encryption Standard (DES) and the Advanced Encryption Standard (AES), by the implementation of a pseudorandom number generator (PRNG). Hence, a brief literature review will be included in this section.

A. The Data Encryption Standard (DES)

The widely used DES crypto-system was first developed by an IBM team and modified by the National Security Agency (NSA) to be adopted by the National Bureau of Standards (NBS) in 1976. It is an iterative block cipher system with a block size of 64 bits. It implements 16 rounds using a 56-bit key that changes for each round according a key generation algorithm. Confusion and diffusion were guaranteed through various substitution and transposition steps [1-4]. It was standardized in 1977 by the National Institute of Standards and Technology, and used internationally since then. It was secure enough at the beginning, however, due to its comparatively small key space, the existence of some weak and semi-weak keys, and the vast increase in the computing power, breaching its security became an easy task.

The DES algorithm weakness and vulnerability was exploited in the last decade of the twentieth century. Electronic Frontier Foundation was able to break DES in 1998 using the so called DES cracker [5]. Around the same period, DESCHEALL project, led by Rocke Verser, Matt Curtin, and Justin Dolske, were also able to break DES. They used idle cycles of thousands of computers across the Internet.

Encryption twice using DES (or 2DES) which doubles the key length to 112 bit was suggested as a modification to DES, but unfortunately, it suffered from man-in-the-middle attack. This drawback lead to a minor improvement in the key space by only increasing the key length from 56 to 57 bits [6].

The drawbacks of 2DES lead to the development of 3DESs which was a far more secure cryptosystem than DES. It was developed by an IBM team in 1999. The application of 3DES with three different keys extends the key space by practically achieving a key length of 168-bit, thus securing the system for few more years to come [7].

Many other variants of DES with less computational efforts were suggested, such as DES-X with key space enhanced by XOR'ing with other elements before and after the encryption process, and GDES that speeds up encryption. However, they were susceptible to differential cryptanalysis [8, 9].

The need arose for a successor to DES, and accordingly, National Institute of Standards and Technology (NIST) put forward a competition for designing a strong encryption algorithm. The criteria for the competing algorithms were to be efficient and easy to implement using both hardware and software, besides being royalty-free in order to be used internationally [10]. This competition was won by the Belgian cryptographers Joan Daemen and Vincent Rijmen in 2000 [11] and was named as the Rijndael algorithm, carrying the acronym of some characters of their names (pronounced "Rhine doll"). Then this algorithm wass termed Advanced Encryption Standard (AES) and defined by FIPS 197. It was approved by the US government to be used for secret and top Secret classified information [12].

B. The Advanced Encryption Standard (AES)

AES is an iterative symmetric cryptosystem operating on 128 bits data block size, i.e. double the data size for DES. There are three variants of AES according to the key lengths; 128, 192 and 256 bits, and the number of rounds; 10, 12, and 14 rounds. These increases in block size, key length and the number of rounds have given the AES algorithm dramatic security improvements as compared to DES when a brute force attack is used, besides no trace of "weak" and "semi-weak" keys are detected so far.

Diffusion and confusion are achieved in the AES computation through four operations that are executed in every round. Those operations are: byte substitution, shift rows, mix columns and add round keys. Also, an excellent key generation algorithm is implemented to produce a different key for each round. It implements S-box tables resulting from transformation using the Galois Field GF(2^8). It defines the transformation algebraically using the GF(2^8) field with the irreducible polynomials ($x^8 + x^4 + x^3 + x + 1$) [9, 10]. The detailed design and operation of the AES algorithm will not be listed here but can be found in the literature [9–12].

Although potential attacks against the AES algorithm, such as interpolation, saturation, Gilbert-Minier, truncated differential, and related-key attacks were suggested by Rijndael, most attacks have focused on the "side-channels", which rely on weaknesses in the security of the application rather than the algorithm [11]. Besides the strength of its security, AES can efficiently be implemented in both hardware and software, which makes it safe and practically beneficial now and for years to come.

C. Pseudo Random Number Generators (PRNGs)

Deterministic or Pseudo-random number generators are algorithms used to generate sequences of numbers having an approximate random property [13]. Pseudo-random number generation is initiated using relatively small key seeds, and the numbers are easy to generate and reproduce. PRNGs are classified into integer generators, sequence generators, integer set generators, narrators, sequence generators, integer set generators, Gaussian generators, decimal fraction generators or row random byte generators. This classification is based on the type of data they produce, such as integers, integer sequences, sets of random integers, integers that fit normal distribution or numbers in the 0 and 1 range with configurable decimal places, respectively. Each of the mentioned types is useful for many cryptographic purposes [14].

Some PRNGs generate pseudo-random numbers using seeds supplied by chaotic systems (dynamic, iterative, decimation) to achieve high speed and good security [15-17]. They are advantageous in having unpredictability or disorder-like, that are required for generating complex sequences. However, they have the problems of non-ideal distribution and short cycle length.

Behnia et. al. proposed a cryptographically secure algorithm for the generation of PRNGs based on three coupled and mutually perturbed Lagged Fibonacci generators [18]. It includes bitwise XOR cross-addition of each generator output with the right-shifted output of the nearby generator. It showed enhanced entropy and acceptable repetition period than the conventional Lagged Fibonacci Generator.

An enhancement to the work discussed above was done through a multi-stage PRNG algorithm that is based on Shannon's concept of confusion and diffusion. This algorithm was designed and tested for randomness using NIST randomness tests by the authors [19, 20]. It implements bitwise manipulation in order to achieve adequate bit string confusion and diffusion by combining various processes such as bit swapping, modular operations and secret splitting techniques. This algorithm will be implemented in this work to improve the key generation and manipulation of symmetric cryptosystems, such as DES and AES.

III. The Multi-Threaded Block Encryption Scheme

This paper proposes a multi-threaded block encryption scheme (MTBES). It is designed with the notion to enhance symmetric cryptosystems by improving the diffusion and confusion processes. This will be done through the introduction of more randomness into the key generation algorithms and utilizing the multi-threaded features in modern computers. In this work, the two widely used cryptosystems, namely DES and AES will considered. Each of these systems includes a number of rounds, where each round requires a certain sub-key. These sub-keys are normally generated by an algorithm that starts with an input secret key. Basically, this research work suggests two modifications; first, the incorporation of a pseudo-random number generator that participates in the generation of the rounds sub-keys needed for either DES or AES cryptosystems. Second, splitting the original message into sets of packets through various threads in the processer, that will be encrypted concurrently, each thread uses the suitable PRNG sub-keys, and then in the end, they are mixed and transmitted to the recipient where the packets are sorted and then decrypted. These two modifications are described in the following sections.

A. Sub-key Generation

Generally, each cryptosystem requires a secret key of certain length, namely it is of 64 bits length for DES and 128, 192, or 256 bits for AES. This key is normally used to generate a set of sub-keys K = {K_1, K_2, …, K_n} according to fixed procedure, where n is the number sub-keys required by the system. The number of sub-keys depends on the system used, namely 16 sub-keys for DES and 10, 12, or 14 sub-keys for AES different key length 128, 192 0r 256 bits, respectively (as shown in fig 1 for DES for example).

In this paper, a PRNG is used to randomly generate another set of n sub-keys, S = {S_1, S_2, …, S_n}. To generate this set of sub-keys, PRNG requires a secret key, too to be used as seed. Each of these sub-keys length is the same as that for K and S. Next, the generated sub-keys, K and S are XOR'ed with each other producing a set of sub-keys K_i as illustrated in Figure I.

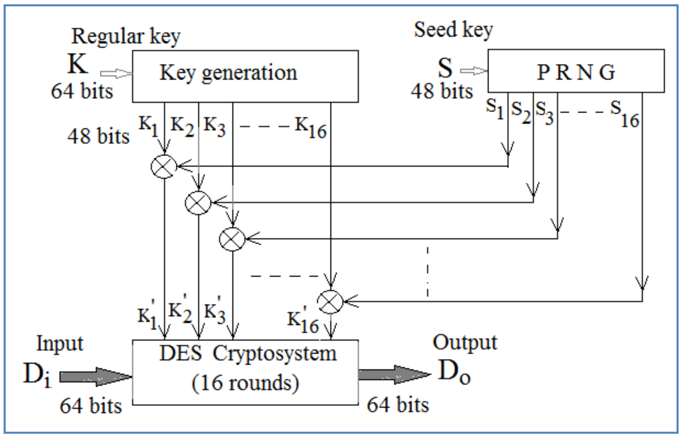


Fig. 1. Block diagram for the proposed sub-key generation scheme

This resulting set of sub-keys will be the one used for the successive rounds of the system under consideration.

As an example, the PRNG implemented in this work that combines logical operation and bits manipulation to achieve the confusion and diffusion concept. It accepts a certain secret key (as a seed) of the required length consisting of any alphanumeric and special characters agreed upon by the communicating parties. The produced sub-keys lengths and the secret seed length depend on the cryptosystem under consideration, (for example, 48 bits for DES and 128, 196, or 256 bits depending on the AES type used). This PRNG is designed and tested for accepted randomness using NIST randomness criterion [20].

B. Multi-threaded Operation

A program is written to arrange the algorithm execution through a multi-threaded processor, which means that its operation is divided over a number of threads. The number of threads is determined by the size of the data to be encrypted, as each thread should be responsible for encrypting a piece of the original text. The number of threads is determined by reading the threading capability of the CPU from the OS and segment the data to fit such threading capability. This process has exhibited an efficient execution practice that is expected to enhance the time complexity measurement.

Multi-threaded programming is used to enhance the performance of the algorithm when it is executed on a computer supporting multi-threading. However, the algorithm operates perfectly on a single core processor that does not support multi-threading.

The algorithm utilizes multi-threading by splitting the data to be encrypted into a number of packet lists equal to the possible number of threads supported by the processor. Each packet of each packet list is then encrypted using a separate thread, and then added to a master packet array in their original order. This order is preserved regardless of unpredictability of thread execution behavior as a packet is placed in the correct location within the array. This master array of packets is then converted to a string representing the final encrypted message, which is finally sent to the recipient.

IV. EXPERIMENTAL RESULTS

The proposed algorithm is incorporated in both DES and AES cryptosystems in order to change them to modified versions, named randomized key DES (named RKDES) and randomized key AES (named RKAES).

The criteria used for the test is the average avalanche effect. The avalanche phenomena may be defined as the percentage of change in the ciphertext contents when the input plaintext is altered. The resulting Average Avalanche Effect percentage (AAE) for these algorithms are compared with original DES and AES cryptosystems running on the same computing environment. Moreover, different input plaintext lengths were considered ranging from 512 bits to 1048576 bits with various number of iterations ranging from 100 to 10000 epochs. These experiments were repeated for three different combinations of input data, namely, numeric only, alphanumeric and Unicode. In the following, some selected results are displayed.

The average avalanche effect (AAE) percentage for the original AES and the modified AES with random key RKAES are calculated for different data sizes ranging from 512 to 1048576 bits, and for different numbers of iterations ranging from 100 to 10000 iterations. The obtained results for the case of 10000 iterations are listed in tables I, II, and III. A graphical representation of the data is shown in the figures II, III, and IV.

TABLE I. AVERAGE AVALANCHE EFFECT OF AES AND RKAES FOR NUMERIC DATA AFTER 10000 ITERATIONS

Average Avalanche Effect of AES and RKAES for Numeric data after 10000 iterations		
Data size	*AAE AES*	*AAE RKAES*
512	36.50%	47.30%
4096	37.00%	47.60%
65536	37.10%	47.90%
1048576	37.60%	52.90%
Average	**37.05%**	**48.93%**

TABLE II. AVERAGE AVALANCHE EFFECT OF AES AND RKAES FOR ALPHANUMERIC DATA AFTER 10000 ITERATIONS

Average Avalanche Effect of AES and RKAES for Alpha-numeric data after 10000 iterations		
Data size	*AAE AES*	*AAE RKAES*
512	36.80%	47.60%
4096	36.80%	47.70%
65536	36.90%	47.90%
1048576	36.80%	53.80%
Average	**36.83%**	**49.25%**

TABLE III. AVERAGE AVALANCHE EFFECT OF AES AND RKAES FOR UNICODE DATA AFTER 10000 ITERATIONS

Average Avalanche Effect of AES and RKAES for Unicode data after 10000 iterations		
Data size	*AAE AES*	*AAE RKAES*
512	36.50%	48.40%
4096	36.70%	48.50%
65536	36.80%	48.80%
1048576	36.70%	54.20%
Average	**36.68%**	**49.98%**

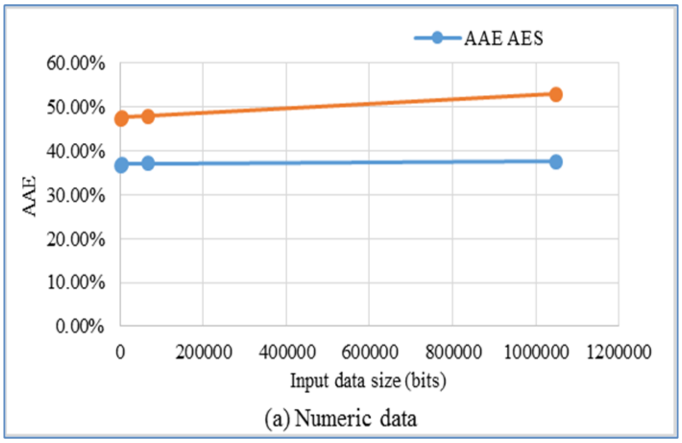


Fig. 2. Average Avalanche Effect of AES and RKAES for Numeric data after 10000 iterations

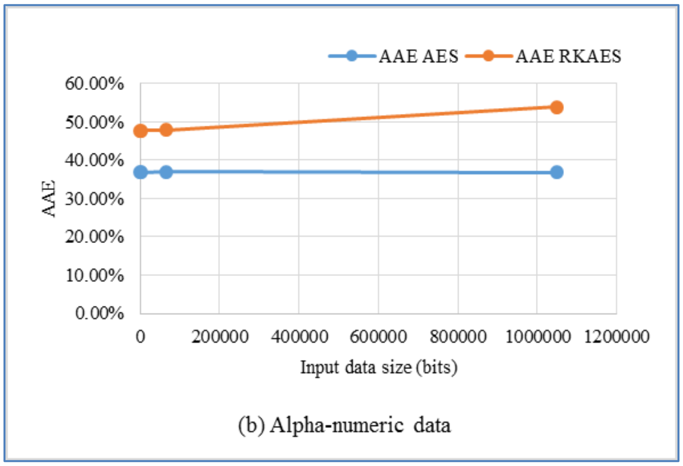


Fig. 3. Average Avalanche Effect of AES and RKAES for Alphanumeric data after 10000 iterations

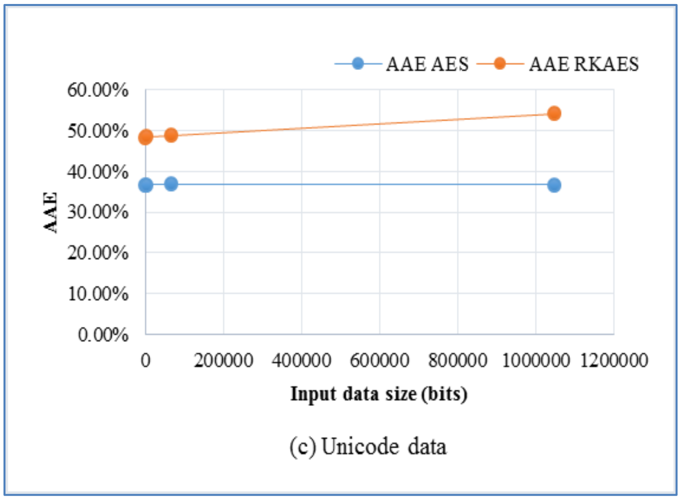


Fig. 4. Average Avalanche Effect of AES and RKAES for Unicode data after 10000 iterations

Similarly, the average avalanche effect (AAE) percentage for the original DES and the modified DES with random key RKDES are calculated for different data sizes ranging and for different numbers of iterations as those for the AES cryptosystem and the obtained results for the case of 10000 iterations are listed in tables IV, V, and VI, and illustrated in the figures V, VI, and VII for the three types of data.

TABLE IV. AVERAGE AVALANCHE EFFECT OF DES AND RKDES FOR NUMERIC DATA AFTER 10000 ITERATIONS

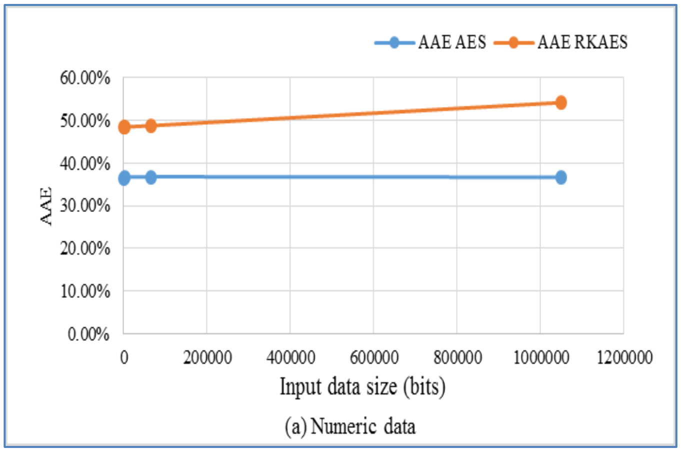
Average Avalanche Effect of DES and RKDES for Numeric data after 10000 iterations		
Data size	*AAE AES*	*AAE RKAES*
512	23.90%	37.50%
4096	23.80%	37.40%
65536	23.70%	37.60%
1048576	23.80%	38.10%
Average	**23.80%**	**37.65%**

TABLE V. AVERAGE AVALANCHE EFFECT OF DES AND RKDES FOR ALPHANUMERIC DATA AFTER 10000 ITERATIONS

Average Avalanche Effect of DES and RKDES for Alpha-numeric data after 10000 iterations		
Data size	*AAE AES*	*AAE RKAES*
512	24.20%	38.00%
4096	24.40%	38.10%
65536	24.70%	38.70%
1048576	24.80%	39.40%
Average	**24.53%**	**38.55%**

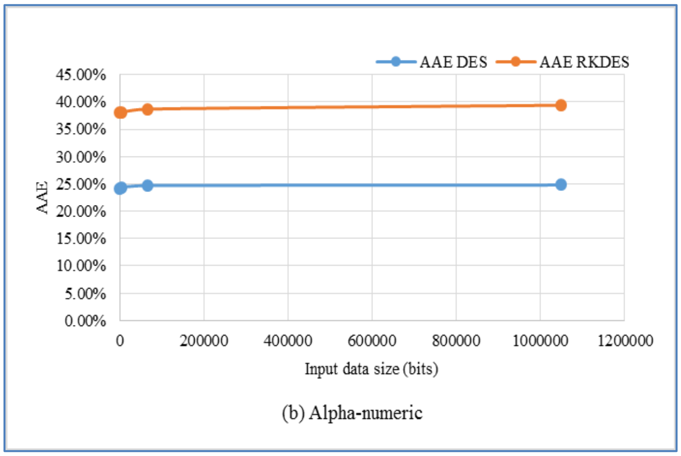
TABLE VI. AVERAGE AVALANCHE EFFECT OF DES AND RKDES FOR UNICODE DATA AFTER 10000 ITERATIONS

Average Avalanche Effect of DES and RKDES for Unicode data after 10000 iterations		
Data size	*AAE AES*	*AAE RKAES*
512	24.50%	39.00%
4096	24.40%	39.00%
65536	24.90%	39.70%
1048576	25.10%	40.60%
Average	**24.73%**	**39.58%**



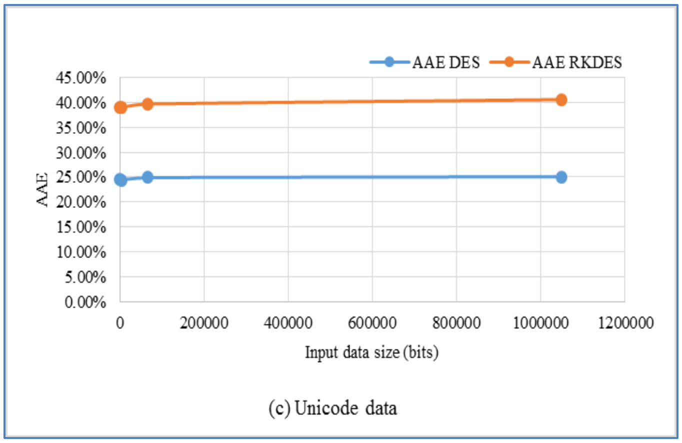
(a) Numeric data

Fig. 5. Average Avalanche Effect of DES and RKDES for Numeric data after 10000 iterations



(b) Alpha-numeric

Fig. 6. Average Avalan*che Effect of DES and RKDES for* Alphanumeric data after 10000 iterations



(c) Unicode data

Fig. 7. Average Avalanche Effect of DES and RKDES for Unicode data after 10000 iterations

The average improvement of the avalanche effect when the key is randomized by the incorporation of the PRNG in the sub-key generation can be calculated by the formula shown in equation (1).

Average improvement of the avalanche effect,

$$\zeta = \frac{\Delta\ AAE}{AAE}\ \% \qquad . \ . \ . \ . \qquad (1)$$

Computing the average improvement of the avalanche effect, ζ for various combinations of data types, input sizes, and number of iterations performed produced the results listed in table III.

TABLE VII. AVERAGE IMPROVEMENT OF THE AVALANCHE EFFECT ζ

Data type	No. of iterations	Average Improvement, ζ (%)	
		DES	AES
Numeric	100	32.53	32.7
	1000	32.05	32
	10000	32.06	32.1
Alphanumeric	100	33.19	33.2
	1000	34.16	34.2
	10000	33.75	33.7
Unicode	100	32.22	32.2
	1000	35.29	35.3
	10000	36.26	36.3
Average Improvement, ζ		33.50	33.52

Table III indicates a considerable improvement in cryptographic strength or system security. The calculated average avalanche effect showed and improvement of more than 33%. Actually the average improvement when various parameters are considered for DES was 33.50% and for AES was 33.52 %, which are almost equal. Such improvement has resulted from the involvement of the PRNG involvement in generating the sub-keys, which indicates that such technique would prove useful in other block cipher systems.

V. RESULTS ANALYSIS

Application of the proposed PRNG algorithm modification as part of the sub-key generation process within AES and DES algorithms has resulted into considerable improvements in the diffusion attribute of both algorithms. This was observed through the considerable increase in the avalanche effect (AAF), which was measured using a custom software package, developed for the testing of encryption strength attributes of block ciphers. The avalanche effect measurements for DES showed an increase by 33.5% in the case of the modified algorithm compared to the original DES, while that for AES showed an enhancement of 33.52% in the avalanche effect in the case of the modified algorithm compared to the original AES.

It can also be stated that the incorporation of PRNG as part of the sub-key generation process, can be considered a form of cryptography applied on the original key and subsequent sub-keys, in a cascading manner. This leads to what is known as domino effect that enhances the confusion and diffusion attributes for block ciphers by applying a multi-stage sub-key generation process.

Moreover, the incorporation of a key encryption algorithm exhibiting highly random outcomes, as measured by the NIST pseudo-randomness tests, such as the implemented PRNG in this work, would lead to enhanced bit diffusion behavior within multi-stage, multi-sub-key block ciphers such as DES and AES which were considered here.

From tables VII, it can be observed that the average avalanche effect for data constructed from larger alphabets was greater than that observed for data constructed from smaller alphabets. Namely, data in Unicode format showed a larger enhancement in the average avalanche effect than alpha-numeric data, and numeric data for the same data size and number of iterations involved. This is also manifested when comparing the average avalanche effects for numeric and alpha-numeric data, i.e. alpha-numeric data, shows better enhancement than numeric data. Besides, when different data sizes are compared, larger data samples showed better enhancement in the average avalanche effect than smaller data samples.

VI. CONCLUSIONS

Significant improvement has clearly resulted due to the incorporation of pseudo-random number generation into the sub-key generation process for both DES and AES algorithms. This means that such a process significantly enhances the diffusion property of the algorithm. This in turn has led to a higher level of security than those obtained using the original algorithms. Moreover, it was noticed that the average avalanche effects get better as one goes from numeric to Unicode through alphanumeric data with increasing number of iterations.

Furthermore, security is achieved by splitting the original message into packets, where each set of packets is encrypted using a pseudo-random sub-key. Using different sub-keys for encrypting sets of packets increases the difficulty of cryptanalysis through differential attacks which require the presence of a large number of original messages and their corresponding cipher texts.

REFERENCES

[1] Bruce Schneier, 1996, “Applied Cryptography: protocols, algorithms and source code in C”, John Wiley & Sons.

[2] William Stallings & Lawrie Brown, 2015, “Computer Security: Principles and Practice”. 3rd Ed., Pearson Press.

[3] M. Ebrahim , S. Khan, and U. Bin Khalid, “Symmetric Algorithm Survey: A Comparative Analysis”, International Journal of Computer Applications, Vol. 61, No.20, January 2013

[4] FIBS, FIPS PUB 46-3 FEDERAL INFORMATION PROCESSING STANDARDS PUBLICATION, 1999. http://csrc.nist.gov/publications/fips/fips46-3/fips46-3.pdf

[5] Curtin, M and J. Dolske, “A Brute-Force Search of DES Keyspace”, Login: The Usenix Magazine, Vol. 23, No. 3, May 1998.

[6] Andrew D., K., “Computer Security”, Michaelmas, Oxford University, 2014. http://www.cs.ox.ac.uk/andrew.ker/docs/computersecurity-lecture-notes-mt2014.pdf.

[7] Noura Aleisa, “A Comparison of the 3DES and AES Encryption Standards”, International Journal of Security and Its Applications Vol.9, No.7, 2015, pp.241-246.

[8] Eli Biham and Adi Shamir, “Differential Cryptanalysis of the Data Encryption Standard”, Springer New York, Nov 9, 2011.

[9] William Stallings, “Cryptography and Network Security: Principles and Practice”, Pearson Education, Prentice Hall, Feb 18, 2016.

[10] Behrouz Forouzan, “ Cryptography and Network Security”, McGraw-Hill, 2008.

[11] Joan Daernen · Vincent Rijrnen, “The Design of Rijndael AES-The Advanced Encryption Standard” https://autonome-antifa.org/IMG/pdf/Rijndael.pdf

[12] Federal Information Processing Standards Publication 197, “Announcing the Advanced Encryption Standard (AES), 2001.http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

[13] D. Dilli, and S. Madhu, “Design of a New CryptographyAlgorithm using Reseeding -Mixing Pseudo Random Number Generator,” IJITEE, vol. 52, no. 5, 2013.

[14] J. M. Bahi, and C. Guyeux, “Topological chaos and chaotic iterations, application to hash functions,” IEEE World Congress on Computational Intelligence WCCI’, Barcelona, Spain, July 2010. Best paper award, PP 1–7,

[15] J. Bahi, C. Guyeux, and Q. Wang, “A novel pseudo-random generator based on discrete chaotic iterations,” INTERNET’09, 1-st International conference on Evolving Internet, Cannes, France, August 2009, PP 71–76.

[16] J. Bahi, C. Guyeux, and Qianxue Wang, ”A pseudo random numbers generator based on chaotic iterations; Application to watermarking,” International conference on Web Information Systems and Mining, WISM 2010, vol. 6318 of LNCS, Sanya, China, October 2010, PP 202–211.

[17] Y. Hu, X. Liao, K. W. Wong, and Qing Zhou, “A true random number generator based on mouse movement and chaotic cryptography,” Chaos, Solitons & Fractals, vol.40, no. 5, 2009, PP 2286–2293.

[18] S. Behnia, A. Akhavan, A. Akhshani, and A.Samsudin, “A novel dynamic model of pseudo random number generator,” Journal of computational and Applied Mathematics –Journal of Computer and Appl. Math, vol. 235, no. 12, 2011, PP 3455-3463.

[19] Adi A. Maaita, Hamza A. A. Al_Sewadi, Abdulameer K. Husain, and Osama M. Al-haj, “A cryptographically secure Multi-stage pseudo-random number generator”, International Journal of Applied Research in Computer and Communication Engineering IJARCCE, Vol. 4, Issue 5, May 2015, DOI 10.17148/IJARCCE.2015.4503, PP 12-18.

[20] Adi A. Maaita, Hamza A. A. Al_Sewadi, “Deterministic Random Number Generator Algorithm for Cryptosystem Keys”, World Academy of Science, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:4, 2015, PP 972-977.

Review of Image Compression and Encryption Techniques

Emy Setyaningsih
Doctoral Program Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia
Department of Computer System, Institut Sains dan Teknologi AKPRIND Yogyakarta, Yogyakarta, Indonesia

Retantyo Wardoyo
Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta,
Indonesia

***Abstract*—In line with a growing need for data and information transmission in a safe and quick manner, researches on image protection and security through a combination of cryptographic and compression techniques begin to take form. The combination of these two methods may include into three categories based on their process sequences. The first category, i.e. cryptographic technique followed by compression method, focuses more on image security than the reduction of a size of data. The second combination, compression technique followed by the cryptographic method, has an advantage where the compression technique can be lossy, lossless, or combination of both. The third category, i.e. compression and cryptographic technologies in a single process either partially or in the form of compressive sensing(CS) provides a good data safety assurance with such a low computational complexity that it is eligible for enhancing the efficiency and security of data/information transmission.**

Keywords—cryptography; compression; lossless; lossy; compressive sensing

I. Introduction

The development of informational technology has a broad impact on the human ways of communication from initially through conventional means to digital ones. Communication through messaging service has also evolved from SMS (Short Message Service) to MMS (Multimedia Messaging Service). Messaging transmission service through internet media such as e-mail, and social media like Twitter, WhatsApp, Facebook, BBM, etc., can also be done.

One emerging problem is that a growing size of digital data, particularly still images, is inevitable due to the need of high-quality images. As a result, a need for larger storage spaces follows. Although storage techniques in digital computers have experienced rapid development, in many situations they require the reduction of digital data storage. One such reduction manifests in the form of bandwidth limitation in communication systems to provide a faster data transmission through communication lines and a smaller percentage of download and upload failure[1]. In addition to the speed of data exchange of a growing size, data safety is of utmost concern due to the susceptibility of data sent through communication lines to their being stolen or extracted by eavesdroppers.

In theory, compression and cryptography are two opposing techniques. Encryption ensures that transmitted data is reliable and integral by converting it from legible into illegible data through an encoding process. Conversely, a compression method seeks to reduce the size of transferred or stored data by finding out and removing duplicate parts of evidence or patterns of data[2]. However, data compression and cryptographic system are deeply connected and mutually useful that they are capable of being employed together. The aims are to generate a smaller size of data; to ensure a quality of data during reconstruction; to speed up data transmission; to reduce bandwidth requirement, and to ensure its safety[3].

In this paper, the author will mainly discuss a combination of compression and cryptography techniques to enhance efficiency in the transmission and safety of image data during the last decade.

II. The Procedure of Sorting Out Literature

In line with a growing need for data and information transmission in a safe and quick manner, researches on image protection through a combination of cryptographic and compression techniques begin to take form. Combination of these two methods may be classified into three categories based on their processual sequences: (1) a cryptographic technique followed by a compression technique [encryption-compresssion], (2) a compression technique followed by a cryptographic method [compression-encryption], and (3) both techniques employed in a single process [hybrid compression-encryption].

The procedure type of literary works is done by seeking out articles in journals and conference proceedings, published from 2004 up to 2016. This searching uses ontology of hybrid image compression encryption mapped and taken from several sources: IEEEXplore Digital Library(IEEEXplore), Science Direct(Direct), Springer, Scholar and other journals and proceedings outside IEEEXplore, Direct, Springer, Scholar, and others. This procedure results in 64 articles with the following details: IEEEXplore (10 articles), Direct (11 articles), Springer (17 articles), Scholar (20 articles), and others (6 articles). Step two: 64 articles is classified into 3 (three) based on their techniques: compression-encryption, encryption-compression, and hybrid compression encryption. Classification of those articles results in 47 (73.44%) relevant articles as shown in Fig 1.

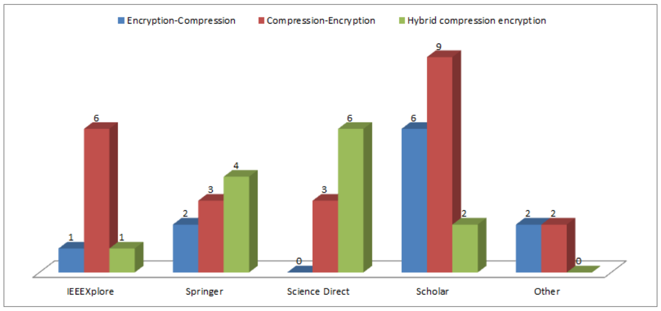


Fig. 1. Articles sorted by classification

Analytical result of 47 articles can be classified into three groups as shown in Fig 2. There are 11 articles (23.40%) in the First group discussing the development of cryptographic techniques followed by compression techniques. The second group of 23 articles (48.94%) discusses the development of compression techniques followed by cryptographic techniques. The third group of 13 articles (27.66%) presents the combination of both techniques.

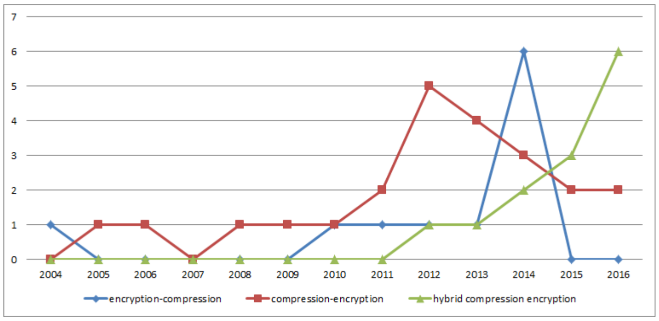


Fig. 2. Research developments according to yearly classification

III. Encryption-Compression Technique

A. *Symmetric Cryptography Method with Lossless Compression*

Johnson et al.[4] and Liu et al.[5] used the combination of a symmetric cryptographic technique using stream cipher method followed by a lossless compression technique using Slepian-Wolf coding. Johnson et al.[4] used a Pseudo-Random Key Generator (PRG), whereas Liu et al.[5] proposed an efficient way of compressing encrypted images through resolution progressive compression (RPC) to avoid exploiting Markov properties in Slepian-Wolf decoding to reduce the complexities of a decoder significantly. In this method, incompatible pixels for encoder are re-correlated to make them closer to a decoder to generate access to low-resolution images. The testing result of entropy value shows that this method has a much better coding efficiency and less computational complexity. Mariselvi and Kumar[6] has also proposed the compression of encrypted images through RPC. The symmetric cryptographic employed is DES algorithm followed by lossless compression technique using Huffman coding or arithmetic coding. The colored images of encryption using DES algorithm are subsequently downsampled to generate sub-images. Each sub-image is then encoded using Huffman or arithmetic coder for performance comparison. Testing of the proposed method is done at four grayscale images to measure Peak Signal Noise Ratio (PSNR) and Compression Ratio (CR) when using arithmetic coding and Huffman coding. The testing result of PSNR values and their compression ratios indicates that Huffman Coding generates higher scores than those of arithmetic coder.

Sharma et al.[7] conducted researches by combining symmetric cryptographic technique using 2D methods Fractional Multiple Parameter Discrete Fourier Transform (MPDFRFT) followed by lossless compression method using zig-zag, Run Length methods, and Huffman encoding. The proposed scheme provides two freeways of data encryption and compression. The test is applied to 3 grayscale images and five colored images and shows a significant increase in their PSNR values. The highest PSNR values of Lena, a cameraman, a baboon, and a satellite image are 76.4, 74.1, 80 and 79.8 dB respectively, with their CR scores are 20%. The lowest PSNR values of each image are 39.8, 34.8, 36.1 and 23.2 dB and their CR is 70%. The proposed scheme also shows a high resistance to brute-force attack seen from the analysis of visual image that looks random cipher. It also provides astounding features in terms of time needed to execute the algorithm and of high sensitivity to the original key.

B. *Symmetric and Asymmetric Cryptographic Method with Lossless Compression*

Shafinah and Ikram[8] have applied a concept of Pretty Good Privacy (PGP) developed by Phil R. Zimmermann to enhance digital file safety for textual data. PGP concept of merging technique is applied using IDEA symmetric cryptography and asymmetric RSA method, with lossless compression technique using ZIP. By contrast, Kale et al. [9] combine symmetric cryptographic techniques 3D-Advanced Encryption Standard (3D-AES) and asymmetric cryptography using the RSA method, with lossless compression technique using Shanon fano. The method of 3D-AES is used to generate symmetric keys by randomizing first key arrays three times which generates a better key in each randomization. As a result, the final key will be stronger than standard AES keys. This technique is capable of providing a high level of informational protection of message confidentiality, and originality exchanged between two parties as well as reducing the length of words. This application works on smartphones and does not require other encryption tools. In contrast, Arunkumar and Prabu[10] proposed the combination of an asymmetric cryptographic technique using RSA method and lossless compression technique using SPIHT method. This combination method allows a partial data access on the part of decoder so that it produces a better efficiency and less computational complexities than the existing approaches. Hence, it will likely be a prospective avenue for video compression in the future.

C. *Symmetric Cryptographic Method with Lossy Compression*

Some researchers have developed a conventional technique of symmetric key cryptographic method and lossy compression method [11]-[14]. Razzaque and Thakur [11] used an image compression method to minimize bit numbers in the post-encrypted images to protect them against unauthorized access. The encryption processes images without the secret key exchange process. You do this by dividing the test images into four blocks, then performing the image encryption using sender's K1 key on one of the essential image blocks and then delivered a receiver it. The receiver then encodes the image he/she received using his/her K2 key and sends it back to the sender. Subsequently, the sender decrypts it using a K1 key and then compresses it using Discrete Cosine Transform (DCT) and sends it back to the receiver. The testing result of 5 grayscale images of 512x512 size with ratio 8 indicates that average PSNR value is 26.35 dB. This size means a still relatively good quality of images after their network transmission.

Kang et al.[12] proposed the application of lossy scalable compression technique after cryptographic process using standard stream cipher method. The values of image pixels that have been encrypted with standard a stream cipher are then put into the compression process by sending subsamples and bit planes. This proposed scheme has an advantage on the part of the decoder as there are no intensive computational iterations and no other orthogonal matrices. It is also applicable to soft and rich-in-texture images. The testing result of 4 grayscale images gives average PSNR values of over 30 dB, indicating that the quality of image remains fairly good. This method is also resistant to statistical attacks as is randomly observed from the visual test.

Aujla and Sharma[13] proposed a combination method of the symmetric cryptographic technique using random permutation method and lossy compression technique using Haar and Daubechies wavelet transformation method to enhance the efficiency of compression process of already encrypted images. The application of this approach results in a positional change for the similar pixel values after their encryption. The resultant images are almost identical to the original as the correlative values among neighboring pixels are relatively high. The result of the encrypted image compression, using orthogonal wavelet transform, is that the majority of the pixels is converted into a series of coefficients. There will be a reduction of data if you remove redundant information contained in the coefficient. This application of compression approach to encrypted images proved to be more efficient according to a testing on CR, Mean Square Error (MSE), and PSNR.

Kamble and Manwade[14] proposed a symmetric cryptographic technique on colored images using Blowfish and block cipher methods followed by a lossless compression method using LBG (Linde-Buzo-Gray) vector quantization algorithm. A test on 6 data samples indicates that the application of a symmetric key algorithm using block cipher and Blowfish methods to encrypt individual colored images requires an average encryption speed of 10.167 byte/second. The quality of encoded images is relatively good, which is over 30 dB.

TABLE I presents a summary to encryption-compression technique reviewed in this section.

TABLE I. ENCRYPTION-COMPRESSION TECHNIQUE SUMMARY

No.	Author, Year	Compression		Cryptographic		Key Stream Generator	Compression Method	Cryptographic Method
		Lossy	Lossless	Symmetric	Asymmetric			
1	Johnson et al.[4], 2004		X	X		PRG	Slepian-Wolf Coding	Binary Stream Cipher
2.	Liu et al.[5], 2010		X	X		-	RPC (Slepian-Wolf Coding)	Stream Cipher
3.	Mariselvi and Kumar[6], 2014		X	X		-	RPC (Huffman or Arithmetic Coding)	DES
4.	Sharma et al.[7], 2014		X	X		-	Zig-zaq scan, Huffman, and RLE	MPDFRFT
5	Shafinah and Ikram[8], 2011		X	X	X	-	ZIP	RSA, IDEA
6	Kale et al. [9], 2014		X	X	X	-	Shanon Fano	RSA, 3D-AES
7	Arunkumar and Prabu[10], 2014		X		X	-	SPIHT	RSA
8	Razzaque and Thakur [11], 2012	X		X		-	DCT	Multiplicative Cipher
9	Kang et al.[12], 2013	X		X			Lossy Scalable Compression	Stream Cipher
10	Aujla and Sharma[13], 2014	X		X			DWT (Haar and Daubechies)	Random Permutation
11	Kamble and Manwade[14], 2014	X		X			LBG Vector Quantization	Block Cipher and Blowfish

IV. COMPRESSION-ENCRYPTION TECHNIQUE

According to Sandoval and Uribe[2], the application of data compression before its encryption will reduce duplicate parts of data that are prone to cryptanalytic exploitation. Also, data compression can speed up an encryption process, and a decryption process will produce corresponding plaintexts. Sharma and Gandhi [15] also supported the idea. They claim that in as many as 70% of the cases studied, implementing cryptography and then compression is more efficient, because: first, compression techniques can eliminate data redundancy, and will work well if the data is random. Therefore, this method can be carried out first before the encryption process. Second: compression can reduce the effectiveness of some attacks. Compression works to reduce data redundancy, whereas cryptanalysis uses a concept of frequency analysis that relies on repeated/duplicate data findings. As a result, if compression is applied beforehand, it may reduce the effectiveness of cryptanalytic attacks that exploit frequency analysis. Third: brute force attacks will take longer time. Brute force attacks are launched in various ways: decrypting data and checking out if consistent output data exists. If a cracker was seeing a compressed data, then a cracker will have first to decrypt and then decompress it to see whether consistent output data exists. It takes a long time, and if the cracker has no idea or does not suspect the probability of data compression beforehand, cryptanalysis will probably not solve it. Fourth: an intruder lacks ciphertext data to do the analysis. An intruder needs enough data to analyze a ciphertext. The fewer clues about internal conditions of a cipher and its key, the better the method. If the compression technique followed by encryption is done, the resulted plain texts will have fewer data redundancies and are thus capable of blocking cryptanalytic attacks. [15].

A. Lossy Compression Method with Symmetric Cryptography

Loussert et al.[3] proposed an integrative model of lossy compression technique using DCT transformation method with an asymmetric cryptographic technique using bit xor operation with fingerprint as the key. The testing result indicates that transmission time increases and systemic security can be increased using biometric characteristics. In this study, the method is applied to a sample of data, and the result shows that the data is capable of being encoded and re-decrypted.

Krikor et al.[16] proposed a selective encryption method to reduce a computational process on large images. Selective encryption aims at obtaining a quick method by encrypting a small piece of a bit stream. The proposed method is in the form of image decomposition into block 8x8. From its spatial domain, the block is later transformed into frequency domain using DCT. Subsequently, DCT coefficient of high-frequency image blocks is encrypted using Non-Linear Shift Back Register (stream cipher). The proposed algorithm for these encryption purposes uses a key of 6-byte long. The first 4 (four) bytes are used to generate a pseudorandom sequence to encrypt images using a stream cipher, and 2 (two) other bytes are two prime numbers used to create rows and columns to randomize images. Based on visual information of randomly perceived encryption result, this proposed method offers a higher security level than if it encrypts all image data.

Benabdellah et al.[17] recommended a compression technique using Faber-Schauder Multiscale Transform (FMT) method followed by quantization on dominant transformed coefficients. Next, the result is encrypted using DES or AES algorithm. The results show that, when using AES, encryption speed is approximately 1,022 times faster than DES method. Both proposed techniques still demonstrate a good performance. The testing result of a visual image looks random, while on FMT-AES the histograms is a Gaussian function, meaning that it is secure from statistical attacks. The quality of reconstructed images is also excellent which is visible from the average PSNR values of over 30 dB for either FMT-AES or FMT-DES methods.

Samson and Sastry[18] proposed a new approach towards image encryption supported by a lossy compression using multilevel wavelet transformation. First, a 2-D multilevel wavelet transformation is applied to input images and then followed by threshold testing on their decomposed structures to obtain compressed images. In this study, Samson tests the application of 5 wavelet filters, i.e., 'haar' 'bior6.8', 'coif5', 'sym8' to see the effect of wavelet filters on the proposed method. The testing result shows that compression ratio depends on types of image and transformation used. Samson and Sastry[19] also suggest a method of securing data that supports RGB images by combining a compression technique using lifting wavelet transform and predictive coding with an encryption scheme using Secure Advanced Hill Cipher (SAHC), involving a pair of involutory matrices, Mix function and an operation called XOR. The test results visually on two pieces of the color image looks random, so that the proposed method can be used to transmit image data efficiently and securely.

Gupta and Silakari [20] introduced a scheme of chaos-based compression and encryption using a cascading 3D cat map and standard map. As for the session to secure key exchange, the use of Elliptic Curve Cryptography is essential. Before its encryption, the image is first compressed using curvelet transformation to remove redundancy in the colored images for a faster transmission. The testing result shows that average PSNR values are over 30 dB, NPCR is over 99%, UACI is below 33%, and entropic values are 7.99 in average, which are close to 8. This shows that the proposed method provides excellent security and speed as well as a better transmission performance.

Li and Lo[21] suggested a combination of image compression and encryption by controlling encryption parameter. The advantage of this proposed compression and encryption combination lies in its applicability on distorted images and its reversibility even without the encryption key. This method uses a base on the JPEG method, by adding an encryption algorithm into its transformation stage. Image encryption and compression method may be employed simultaneously using DCT transformation and block of the 8x8 pixel. It develops a new orthogonal transformation by introducing sign-flip into butterflies method on the DCT flow-graph structure. One of the alternative ways to use during JPEG transformation is a different orthogonal transformation, which is produced by the sign-flipping strategy. By selecting butterflies method for sign-flip, it is expected to control the visual quality of encrypted images. The testing result of significant key space and encryption space, of security against replacement attack, and of security against statistical model-based attack has demonstrated that the proposed method is capable of securing image data.

B. Lossless Compression Method with Symmetric Cryptography

Chung and Kuo[22] suggested two approaches combine encryption with multimedia compression system, i.e., a modified selective encryption using entropy coder with some statistical models. The proposed method works by changing entropy coders into cipher encryption using some statistical models. The test results showed that compression without sacrificing performance and computation speed, security remains achievable.

Hermassi et al.[23] introduced a new scheme called Chaotic Human Tree (CHT) method using a modification of Huffman code implemented on textual data. This approach has succeeded in overcoming the downsides of Multiple Huffman Coding (MHT) by combining stream cipher algorithm and Huffman compression algorithm. By contrast, the cryptographic method used is a chaotic map to generate keystream by renewing Huffman coding tree. Keystream generated is based on the concept of chaos; the permutation is then performed on the base tree without changing their statistical models. As a result, a symbol can be encoded by more than one codeword for data with the same length. An analysis of compression performance results in an exactly same ratio between proposed method and standard Huffman scheme. This fact is, in fact, a consequence as there is no statistical change in the model during Huffman tree mutation. Each symbol encrypted using the proposed method will have the same code length of the code used in the classic Huffman scheme. The proposed method is relatively immune to brute force attacks. In comparison to arithmetic coding, the proposed method has a little higher compression efficiency. However, it has a slower encryption/decryption speed than that of Huffman+stream cipher algorithm. Chen et al.[24] also proposed a scheme of compression and encryption based on chaos. For encryption, they use a table dynamically modified in its searching process. As a result, the target symbol will finally connect to other partitions that result in fewer iterations to find it. Simulations show that the proposed modification offers a better compression performance, while execution efficiency is proportional to its security level.

Kishore et al.[25] proposed the application of Slepian-Wolf coding compression method, while the cryptography is done using bit-wise exclusive OR operation. The study focuses on the design and analysis of lossless compression, where image data is encrypted using stream cipher method after its compression. The proposed method is tested on two grayscale images to check the randomly perceived cipher image visually. The success of this approach lies in its provision of partial access to the source of data on the part of the decoder to increase security.

V. Nair et al.[26] proved that arithmetic coding is randomly not secure. Therefore, a lossless compression method is presented using arithmetic coding technique by dividing data into similar intervals and followed by symmetric encryption technique using bit-wise XOR with pseudorandom bit sequence. This system offers compression and security and is capable of blocking any attacks launched to obtain information about input or output permutation and information on how to divide intervals. The proposed method is proved to be secure and immune to chosen plaintext attack. Also, it is capable of reducing a delay during data transmission and of increasing data security.

Sudesh et al.[27] proposed the application of adaptive compression to obtain a high compression ratio. An adaptive compression works to reduce the size by analyzing frequencies repeatedly and then retaining them in a dictionary or tabular forms. By contrast, cryptography uses Milline transform approach based on the mathematical transformative operation which makes it perform faster and more efficient. The level of security is obtained through the method of implementation transformation Milline encoding, Whereas coding efficiency will be achieved when you apply adaptive dictionary. The testing result of 6 sample images indicates that the average PSNR values are 32.93 dB.

Xiang et al.[28] proposed a Joint compression and selective encryption based on SPIHT(JCSE-SPIHT), i.e., a compression algorithm and selective encryption based on set partitioning in hierarchical trees (SPIHT), by embedding encryption into SPIHT coding procedure. The basic idea of JCSE-SPIHT method is to perform a fast random insertion(FRI) on the list of insignificant pixels(LIP) and insignificant sets(LIS) on selected numbers of iteration coding of SPIHT. Therefore, selective node randomization of LIP and LIS by FRI is in the first round (r) of iteration, where parameter r is used to control the particular encryption strength. A proper selection of r will generate a good trade-off between security requirement and computational overhead. The testing result indicates that r = 6 is a suitable configuration as the plain image is well protected and requires 1-4% of data to be encrypted. The proposed method generates keystream plain text that is dependent on JCSE-SPIHT compression algorithm that makes it immune to chosen-plaintext attacks.

C. *Combination of Lossy and Lossless Compression Method with Symmetric Cryptography*

Ou et al.[29] developed an ICES (Image Compression Encryption Scheme) model by integrating compression technique using Discrete Wavelet Transform(DWT) transformation method, orthogonal wavelet family type Haar without quantizer and Significance-Linked Connected Component Analysis(SLCCA) encoder proposed by Chai et al.[30]. The cryptographic technique used is AES method. The proposed method allows compressed images to generate a high compression ratio while maintaining security during transmission so that simultaneously can solve the problem of bandwidth and safety. The test results on six image grayscale with different image sizes shows that the reconstructed image is of high quality, and efficient.

Alfalou et al.[31] proposed simultaneous fusion, compression, and encryption of multiple images (SFCE) methods to obtain image compression and encryption simultaneously. The proposed techniques adapt the DCT method, by combining spectral fusion according to DCT properties, particular spectral filtering, and quantization of encoded frequency using select bit number. The study finds that this size of adaptation provides a good trade-off between bandwidth spectral plane and output number of reconstructed images. Improved encryption capabilities are achieved by using biometric locks and by randomly changing the angle of rotation of each block before fusion spectral. The use of the image as the key of real-valued has succeeded in increasing compression level into 50% better than that of the original SFCE method.

The following study uses a modification of chaotic key generator on encryption process. Tong et al.[32] proposed an image compression and encryption scheme based on nearest-neighboring coupled-map lattices(NCML) and Non-uniform Discrete Cosine Transform(NDCT). A new chaotic map is recommended based on Devaney theory, which works as a local map of NCML called system spatiotemporal cross chaotic. This algorithm adopts Huffman coding and NDCT for transforming image data and compressing it. It consists of two steps of the encryption process. Compressed data is divided into blocks and is subsequently permutated and diffused amongst blocks simultaneously. The parameter obtained through system spatiotemporal cross chaotic is used to control NDCT non-uniformity, which plays a significant role in the encryption process. The result of security test indicates that the proposed method offers high speed and safety as well as a good compression effect. This is observable from the average

PSNR values of 6 tested images of over 30 dB, average entropy values of over 7.99 which is close to 8, average NPCR values of over 99%, and average UACI values of over 33%. Besides, the degradation result of the performance of the proposed method is 3.26-9.02% better than that of a typical technique of DCT and Huffman coding followed by AES. Tong et al.[33] also conducted a study to combine lossy compression technique using lifting wavelet transform(LWT) and lossless compression technique using SPIHT coder, followed by cryptosystem symmetric using Chaotic sequence generation. Testing of the proposed method is done using five grayscale image data with a size of 512 x 512 pixels. The measurement result of the change rate of cipher text is about 50% (the change rate is the ratio of the position of the original cipher text and cipher text in which the plaintext is modified). The testing result of changing one bit of bitmap image, on the modification level of cipher stream, ranges between 40-44%, indicating a high sensitivity to plain text. Based on the testing of the key sensitivity of five images, an average value of key sensitivity is more than 49.9%, indicating that algorithm has an excellent key sensitivity. Its compression ratio is about 50% of the original file size. The test results histogram also looks flat; it shows that the frequency of appearance of color in the cipher image looks evenly, so is secure against statistical attack. The entropy value is relatively high as well, i.e., 7.99 in average which is close to 8, meaning that this method is secure from cryptanalytic entropy attack.

Zhiqianga et al.[34] combined JPEG image compression algorithm with a chaotic encryption algorithm. This process can save storage space for images and tight transmission security of pictorial information more efficiently. In contrast, Goel, N et al.[35] combined a lossy compression technique using DCT method with a lossless compression technique using Huffman coding, followed by symmetric cryptosystem technique using Logistic Map method. This paper highlights anything to do with Huffman coding in the view of the proposed image encryption method. Besides, it also presents a snapshot of one logistic map dimension, having been used as pseudorandom numbers. The proposed method is shown to overcome many limitations of dictionary-scrambling-based encryption technique. The testing of the proposed method is excellent when implemented on the low-contrast image, as seen from the high PSNR value. Also, the method has high sensitivity key, and use of the compressibility of the encoder does not result in adverse effects.

Kumar and Vaish [36] proposed a compression-encryption image method to transmit image quickly and securely through the network. The core idea of the proposed method is to select significant and non-significant coefficient in the wavelet domain. These two coefficients will be encrypted using pseudo-random number sequence and permutation on their each coefficient. The proposed method is first to perform a DWT transformation process. Furthermore, do the pseudo random encryption process (PRNG) and then the compression process using the quantization and entropy coding, whereas wavelet sub-bands detail (LH, HL, HH) substitution process is carried out using the k2 key and is subsequently encrypted using coefficients permutation. The next process of image encryption result is compressed using Singular Value Decomposition(SVD) and Huffman code. Seeing that performance of image compression is mostly based on the selected wavelet transformation filter, then the use of different filters like biorthogonal wavelet, Haar, Symlets, Daubechies, Coiflets, etc., is also tested. The test results demonstrate that the use of biorthogonal wavelet filter produces better compression performance. For example, when image Lena is compressed using wavelet biorthogonal on singular values (SVs)=256 and η = 1, the CR value is 0.2883 and PSNR value is 45.66 dB. By contrast, CR values for other wavelets like Symlets, Daubechies, Coiflets, Haar and Discrete Meyer wavelet are 0.2970, 0.2967, 0.2979, 0.3014 and 0.3092 each respectively, while appropriate PSNR values are 45.75 dB, 45.95 dB, 45.04 dB, 42.64 dB, 47.89 dB. Also, the proposed method has an advantage of making use of SVD to obtain a better compression performance while maintaining the desired features of the reconstructed image. The proposed scheme is immune to brute force attacks and proved to be more efficient than that of Zhang and to be better than that of JPEG standard.

D. Joint Method of Lossy or Lossless Compression with Asymmetric Cryptography

Rahmawati et al.[37] combined lossy and lossless compression techniques using DCT, quantization, Huffman coding to obtain a high energy compaction, followed by asymmetric cryptosystem technique using Secure Hash Algorithm-1 (SHA 1) method as its encryption algorithm. Errors in one of the keys will generate an impaired, reconstructed image. The value of compression ratio and PSNR obtained through this algorithm is influenced by the employed quantization matrix. Luminance quantization matrix produces a lower compression ratio than that of chrominance quantization matrix, only that it produces higher PSNR values. The proposed algorithm has a high sensitivity to the use of each of the key. The key sensitivity marks a good encryption performance.

Chal.la et al.[38] proposed a Learning with Errors (LWE) and public-key based compression which is implemented using CNA to reduce a key size. CNA is a new lossless compression algorithm which is practical and has a higher adaptive capability.

TABLE II presents a summary to compression- encryption technique reviewed in this section.

TABLE II. COMPRESSION- ENCRYPTION TECHNIQUE SUMMARY

No.	Author, Year	Compression		Cryptographic		Key Stream Generator	Compression Method	Cryptographic Method
		Lossy	Lossless	Symmetric	Asymmetric			
1	Loussert et al.[3], 2008	X		X		fingerprint	DCT	Bit XOR Operation
2.	Krikor et al.[16], 2009	X		X		Pseudorandom	DCT	Selective Encryption, Bit Stream Cipher
3.	Benabdellah et al.[17], 2011	X		X			FMT	DES or AES
4.	Samson and Sastry[18], 2012	X		X			2-D Multilevel Wavelet Transformation	Permutation
5	Samson and Sastry[19], 2012	X		X			Lifting Wavelet Transform	SAHC
6	Gupta and Silakari [20], 2012	X		X		Cascading 3D Cat Map,Standard Map	Curvelet Transformation	Elliptic Curve
7	Li and Lo[21], 2015	X		X		Random 128-bit Key	JPEG	RC4
8	Chung and Kuo[22], 2005		X	X		Segment Key	Multiple Huffman Tables (MHT) or QM Coder	Stream Cipher
9	Hermassi et al.[23], 2010		X	X		Piecewise Linear Chaotic Map	Renewing Huffman Coding Tree	Stream Cipher
10	Chen et al.[24], 2011		X	X		Chaotic Map	Entropy Coding	Lookup Table
11	Kishore et al.[25], 2012		X	X			Slepian-Wolf Coding	Bit-wise XOR Operation
12	V. Nair et al.[26], 2012		X	X		Pseudorandom Bit	Arithmetic Coding Technique by Dividing Data into Similar Intervals	Bit-wise XOR Operation
13	Sudesh et al.[27], 2014		X	X			Adaptive Compression	Transformation Milline
14	Xiang et al.[28], 2014		X	X			SPIHT	Selective Encryption
15	Ou et al.[29], 2006	X	X	X			DWT, SLCCA	AES
16	Alfalou et al.[31], 2013	X	X	X		Biometric	Combining Spectral Fusion According to DCT Properties	XOR Operation
17	Tong et al.[32], 2013	X	X	X		Spatiotemporal Cross Chaotic System	Huffman Coding and NDCT	Packed into blocks, Permutation Between Blocks and Diffusion in Block
18	Tong et al.[33], 2016	X	X	X		Lorenz map, Henon map, Logistic Map	LWT, SPIHT	Stream Cipher
19	Zhiqianga et al.[34], 2013	X	X	X		Logistic Sequence	JPEG	Chaotic Encryption
20	Goel, N et al.[35], 2014	X	X	X		Logistic Map	DCT, Huffman	Dictionary Scrambling
21	Kumar and Vaish [36], 2017	X	X	X		PRNG	DWT, SVD, Huffman	Stream Cipher
22	Rahmawati et al.[37], 2013	X	X		X		DCT, quantization, Huffman	SHA 1
23	Chal.la et al.[38], 2015		X		X		CNA	LWE and Public Key

V. HYBRID COMPRESSION- ENCRYPTION TECHNIQUE

This technique combined a compression method and cryptography, or vice versa. However, that combination is not worked out in a sequential order.

Al-Maadeed et al.[39] proposed a joint method of a selective encryption of an image and a compression. The basic idea of this proposed algorithm is to demonstrate the effect of the application of several keys to enhance security by increasing the number of external keys in each encryption process. The encryption process uses an encryption algorithm based on chaos conducted on the approximation of the results of the DWT transformation. In contrast, DWT transformation results in a detailed component of the compression process. The encryption process of the proposed method uses a key length of 94 bits. It also conducted a comparison of a key length of 97 bits. The fundamental principle of encryption is to use random numbers dependent on original condition to generate this randomized number sequence. This technique creates a significant reduction in encryption and decryption time. The testing result shows a reduction of encryption time into about 0.218 seconds with one key, 0.453 second with two keys, and 0.5 seconds with three keys. Correlation coefficient value between an original image and an encrypted image decreases when the number of external encryption keys increases. And this Resulted in an increase in security (the more the key, the security of the data to be encrypted is also increasing). Al-Maadeed et al.[39] also show how correlation coefficient changes exponentially when it uses a value different from the controlling parameter. Also, they recommend the use of more than 128 bits external keys to enhance the overall security and also suggest other methods for compression.

Wang et al. [40] proposed a similar technique to that of Al-Maadeed et al.[39], the difference on the Schema Lifting(LS) DCT that is performed on the input image before processing the transformation DWT. Having finished performing the separation of subband approximation (LL) and subband details (LH, HL, HH) through DWT transformation process, encryption and compression are done using a different method. After getting subband LL proceed with the encryption method process using a stream cipher, other subbands are encrypted using a permutation method. By contrast, compression is performed by a third party. Regarding subband LL, the result of encryption is then compressed using lossless compression process (encoding is carried out on each coefficient bit). With subbands LH, HL, and HH, encryption results are then compressed using rate-distortion optimized quantization and is followed by a coding process using an arithmetic coding method. The test results of the proposed method are equivalent to the value of the smallest compression ratio (CR = 4.461) when using filters Bior2.2. By contrast, the best-suited subband level for the proposed scheme is on level 3. Also, the proposed scheme provides a small computation time.

Hassan and Younis[41] offered a combination of lossless compression technique using Quadtree and Huffman coding method and symmetric cryptosystem technique using the partial method where the encrypted data will become a part of compressed data using AES method. The testing result indicates that only 10-25% of the output of Quadtree compression algorithm is encryptable. The testing of the proposed method is performed on a grayscale image of size 256x256. The visual testing of a cipher image looks random. The test results histogram also looks flat; it shows that the method is safe from statistical attack. However, the PSNR is low, i.e., below 30 dB, meaning the quality of the reconstructed image is not reasonably safe.

Xiaoyong et al.[42] combined a compression technique using an algorithm of generalized knight's tour, DCT, Quantization and zigzag scan coder and symmetric cryptosystem technique using non-linear chaotic maps method. In contrast, the encoding procedure uses a nested generalized knight's tour (NGKT) matrix generated scramblingly by Semi Ham algorithm on the bright image. Furthermore, this is to produce a high image compression ratios by utilizing DCT and quantization coding. The diffusion process is subsequently done using encryption parts of DCT coefficient obtained from Chen chaotic map. The evaluation of the proposed scheme is carried out by a series of tests using five grayscale images, and the results show that the proposed scheme has a compression performance and good security. Evaluation is also done using compression Degree (CD) used to reflect the compression performance. After the testing result of 5 data, it turns out that the compression performance of the proposed method is better than that of Zang, Yuen, and Zhou, to which this paper refers. However, it is closer to JPEG algorithm. Analysis of key space shows that computational accuracy of 64-bit double precision numbers is about 10-14. The key space of each chaotic map is 1014, and chaotic key space is 1014 × 1014 × 1014 = 1042 which are bigger than 2100[43] that the proposed scheme is relatively resistant to brute force attacks. The testing of key sensitivity provides a value of > 99%, meaning that the key sensitivity is excellent. The testing of differential attacks shows that NPCR value is over 99% and UACI value is over 33%. It means that the proposed scheme is sensitive to plain image and is capable of blocking differential attacks due to its high NPCR and UACI values. The Robustness analysis shows that an image obtained from a decryption process is still recognizable even though it is not as good as the original. The last test is a Structural Similarity Index Measurement (SSIM) comparing images regarding lighting, contrast, and structure, replacing the application of PSNR method in evaluating the similarity among pictures. The testing result of SSIM of 5 data shows a result that is closer to 0, meaning that the proposed scheme is secured.

Hamdi et al.[44] proposed a method using a more efficient compression technique to generate a high-quality image and little computational complexities. The cryptographic method is confusion and diffusion technique which is integrated and connected to compression chains. The first step is to generate three keys for encryption process using Chirikov Standard Map algorithm. The next step is to perform DWT transformation and is followed by a bit encryption on wavelet coefficient (LL Subband) using the first key, whereas other subbands are undergoing encryption process using the list of LIP and second key. The third step is permutation after SPIHT coding. This stage is to increase the diffusion of the encrypted image. It is to ensure an efficient informational diffusion according to bitwise permutation process. The testing result of the image of a house using level-3 decomposition shows that PSNR value is 39.674, while the image of an airplane using level-2 decomposition shows that PSNR value is 38.013. The average key sensitivity of MAD value for ten tested data images with three different keys is 85.13, which is closer to its ideal value, 85.33 (256/3). By contrast, the average number of pixels change rate (NPCR) of 10 tested images for all stages is 99.55% bigger than the required value of 99%, and the value of Unified average changing intensity (UACI) of 33.59% is larger than the required value of 33%. Thus, the result of differential analysis indicates that the proposed encryption algorithm is very sensitive to small changes in the original images and very resistant to differential attacks.

The following several studies use a concept of compressive sensing(CS) to perform compression and encryption process simultaneously[45]-[52]. Zhou et al.[45] proposed an image encryption-compression hybrid algorithm based on CS and random pixel exchanging, where compression and encryption are done simultaneously. The first divides the image into four blocks for the purpose of compression and encryption. Then an exchange of the pixels that have been randomized to be compressed and encrypted. This method makes use of circulant matrices to develop measurement matrices on CS and to control first line vectors of the circulant matrices using the chaotic system. The proposed algorithm is proved to be secure. The simulation shows that the proposed method provides good security and excellent performance of the compression. It is perceived from the histogram of three original images which is clearly different from each other, whereas the encrypted image has a similar histogram. Huang et al.[46] also proposed a CS-based encryption method combining sampling, compression, and encryption simultaneously. The testing result indicates that the proposed encryption method does not achieve an outstanding randomness, even the diffusion and sensitivity outperform image encryption method performed in parallel. The measurement result shows that the average PSNR values of 5 tested data are over 30 dB, indicating a good reconstruction quality. This method uses the key of 128 bits, meaning that it occupies the main space up to 2128≈3.4028 × 1038. In fact, it provides an adequate security against brute force attacks. It is also indicated by its average entropy values of 7.99, which is close to 8. The histogram looks visually flat, meaning that this method is immune to statistical attack. The average coefficient correlative value of adjacent pixels of 5 images is 0.0024, which is close to 0. Apart from that, the mean value of NPCR and UACI is close to 99.61% and 33.46% respectively, meaning that this method is very sensitive to small change in the key.

Fira[47] proposed a method designed to achieve an efficient compression to save memory space, to reduce transmission time, and to reduce energy consumption. CS algorithm is applied to compress and encrypt ECG signals. This study analyzes the compression obtained through standard wavelet-dictionary, while encryption is used to analyze the effect of its projection matrices.

Zhang et al.[48] designed a simple scheme to simultaneously compress and encrypt an image using random convolution and random subsampling methods based on CS encoding to offset the downsides of double random phase encoding which has no compression capability. Utilization of random methods with an underlying convolution CS inspires this method. In this method a CS using convolution with a random pulse followed by random subsampling. The testing shows that the proposed scheme is relatively immune and is capable of blocking the cropping attacks.

Ahmad et al.[49] proposed a new image encryption scheme based on chaotic maps and orthogonal matrices. In addition to performing encryption for higher security, this method also supports partial encryption for a faster process and a better result. The proposed scheme uses a primary method of new properties of the orthogonal matrices to get a random orthogonal matrix using Gram-Schmidt algorithm, and nonlinear chaotic map to randomize the pixel values of a plain image. The proposed scheme is capable of reconstructing an image, even if it is distorted by AWGN/noise due to its transmission through the network. The experiment and security analysis show that the proposed scheme is relatively secure and robust from channel noise and JPEG compression. The output quality of a decrypted image is fairly good. The highest PSRN value is 40 dB, whereas the average PSNR values of 4 tested images are 31.38 dB. The analysis of average differential attack of 4 tested images gives partial NPCR value=99.1% and UACL=15.38%. This fact indicates that the proposed algorithm is very sensitive to input change, but its security is still lacking, i.e., below 33%. The result of histogram analysis of encryption is close to Gaussian distribution, meaning that encrypted histogram is capable of concealing frequency distribution of plain text images.

Chen et al.[50] proposed an encryption and compression scheme based randomly on Elementary Cellular Automata (ECA) and Kronecker Compressed Sensing (KCS). The first stage: encryption is done using ECA to generate an image uniformity at its sparsity level. Second stage: KCS encryption is performed to encrypt and compress randomized images by measuring matrices with a reduced size conforming to the original image size. The proposed Kronecker Compressed Sensing (KCS) is used to solve high computational complexity and a bigger storage demand due to big matrix size. The experiment indicates that the proposed method based on ECA offers excellent performance in randomizing and enhancing uniformity at its sparsity level. Image encryption and compression based on the application of the method gives a higher level of confidentiality and a good performance of compression and flexibility.

Deng et al.[51] proposed a joint algorithm between 2D CS and Discrete Fractional Random Transform (DFrRT), where compression and encryption can be performed simultaneously with a simple operation and high security. Plain text is expressed in the 2D cosine discrete domain and measured from two orthogonal directions. Furthermore, after encrypting using DFrRT do repeated measurements. This scheme shows a good performance by combining CS capability with simple operation of DFrRT. The testing result indicates that histogram of the reconstructed image takes the form of Gaussian function, meaning that the proposed scheme has a high capability to impede statistical analysis attack. Besides, the simulation shows that the proposed scheme is capable of blocking pixel cropping attack, brute-force attack and is sensitive to key change.

Zhou et al.[52] proposed a method of compression-encryption image scheme based on hyper-chaos system and 2D sensing. The parameters of 2D CS used are: x01 = 0.13, x02 = 0.25, μ = 3.99. The original value of hyper-chaos system is stated as: x0 = 0.3, y0 = 0.4, z0 = 0.5 and h0 = 0.6. The result of simulation shows that the proposed compression-encryption image scheme is effective, robust and secured with a good compression performance. This method is capable of blocking statistical analysis, brute force and noise attacks as the key space used is much bigger. Therefore, this proposed algorithm is useful for reducing the storage size of adequate security.

TABLE III presents a summary to hybrid compression-encryption technique reviewed in this section.

TABLE III. HYBRID COMPRESSION- ENCRYPTION TECHNIQUE SUMMARY

No.	Author, Year	Compression			Cryptographic		Key Stream Generator	Compression Method	Cryptographic Method
		Lossy	Lossless	Compressive Sensing	Symmetric	Asymmetric			
1	Al-Maadeed et al.[39], 2012	X			X		Chaotic Maps	DWT	Selective Encryption
2.	Wang et al. [40], 2015	X	X		X			LS DCT, DWT, Quantization and Adaptive Arithmetic Coding	Selective Encryption (Stream and Permutation Ciphers)
3.	Hassan and Younis[41], 2013		X		X			Quadtree and Huffman Coding	Partial Encryption, AES
4.	Xiaoyong et al.[42], 2016	X	X		X		Non-linear Chaotic Maps	DCT, Quantization, Ziqzaq Scan, Entropy Coding	Selective Encryption , NGKT
5	Hamdi et al.[44], 2017	X	X		X		Chirikov Standard Map	DWT, SPIHT	Confusion and Diffusion Technique Which is Integrated and Connected to Compression Chains
6	Zhou et al.[45], 2014			X	X		Logistic Map	CS	Random Pixel Scrambling
7	Huang et al.[46], 2014			X	X		Spatio temporal Chaos	CS	Including Arnold Scrambling, Mixing, S-box, Block-wise XOR Operation
8	Fira[47], 2015			X	X			CS	Substitutions
9	Zhang et al.[48], 2015			X	X			Random Convolution, Random Subsampling Methods Based on CS Encoding	A Linear Transform Encryption Mode and There Are Two Masks
10	Ahmad et al.[49], 2016			X	X		Nonlinear Chaotic dan Logistic Map	DCT, Matrix Orthogonal (via Gram-Schmidt Process)	Partial Encryption (Block- wise Random Permutation, Diffusion Process)
11	Chen et al.[50], 2016			X	X			KCS	ECA
12	Deng et al.[51], 2016			X	X		Logistic Map	2D CS	DFrRT
13	Zhou et al.[52], 2016			X	X		Hiper-Chaos	2D CS	Cycle Shift Operation

VI. CONCLUSIONS

The most combination of Encryption-Compression technique discussed above uses symmetric cryptographic and lossless compression method. In fact, it shows that the process focuses more on image security than on data size reduction. The application of lossless compression technique is to ensure that all data is reversible and can be reverted to the original while maintaining the high quality of reconstructed images and compression ratio. As such, this concept is most applicable when data accuracy is of paramount importance, such as textual information, biomedical image, and legal data. The majority of the measurement of the quality of the decompression image against the original image, the compression ratio as well as the processing time are used to measure the success of the proposed method, while the measurement results cipher visual image is used to analyze the level of security of some of the proposed method.

The combination of Compression-Encryption technique has some advantages because compression method can be lossy, lossless, or combination of both. In contrast, most cryptographic techniques use symmetric cryptography by developing a chaotic method to generate a symmetric key. As such, this approach applies to data image, either audio or video. Conversely, the proposal to use various chaotic methods aimed at generating a symmetric key to enhancing its security.

The hybrid compression-encryption technique is capable of providing real data security assurance with such a low computational complexity that it is eligible for increasing the efficiency and security of data/information transmission. So the concept qualifies for and could improve transmission efficiency and data security by improving the performance of each compression and cryptographic technique through hybrid concept. This concept is expected to be able to combine excellent properties of lossy and lossless compression techniques and to offset the downside of symmetric and asymmetric cryptographic techniques, particularly about cipher key management, to obtain the much smaller size of data, still good quality of data during reconstruction and security assurance.

REFERENCES

[1] M. Merdiyan and W. Indarto, “Implementasi Algoritma Run Length, Half Byte, dan Huffman untuk Kompresi File,” in *Seminar Nasional Aplikasi Teknologi Informasi 2005 (SNATI 2005)*, 2005, pp. 79–84.

[2] M. M. Sandoval and C. F. Uribe, “A Hardware Architecture for Elliptic Curve Cryptography and Lossless Data Compression,” in *15th International Conference on Electronics, Communications and Computers (CONIELECOMP’05)*, 2005, no. March, pp. 113–118.

[3] A. Loussert, A. Alfalou, R. El Sawda, and A. Alkholidi, “Enhanced System for Image’s Compression and Encryption by Addition of Biometric Characteristics,” *International Journal of Software Engineering and Its Applications.*, vol. 2, no. 2, pp. 111–118, 2008.

[4] M. Johnson, D. Wagner, and K. Ramchandran, “On Compressing Encrypted Data without the Encryption Key,” in *Theory of*

Cryptography, First Theory of Cryptography Conference, TCC 2004, Cambridge, MA, USA, 2004, pp. 491–504.

[5] W. Liu, W. Zeng, L. Dong, and Q. Yao, "Efficient Compression of Encrypted Grayscale Images," *IEEE Transactions on Image Processing.*, vol. 19, no. 4, pp. 1097–1102, Apr. 2010.

[6] C. MariSelvi and A. Kumar, "A Modified Encryption Algorithm for Compression of Color Image," *International Journal of Recent Development in Engineering and Technology*, vol. 2, no. 3, pp. 94–98, 2014.

[7] D. Sharma, R. Saxena, and N. Singh, "Hybrid Encryption-Compression Scheme Based on Multiple Parameter Discrete Fractional Fourier Transform with Eigen Vector Decomposition Algorithm," *International Journal of Computer Network and Information Security.*, vol. 6, no. 10, pp. 1–12, Sep. 2014.

[8] K. Shafinah and M. M. Ikram, "File Security based on Pretty Good Privacy (PGP) File Security based on Pretty Good Privacy (PGP) Concept," *Computer and Information Science.*, vol. 4, no. 4, pp. 10–28, 2011.

[9] N. A. Kale and S. B. Natikar, "Secured Mobile Messaging for Android Application," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, no. 11, pp. 304–311, 2014.

[10] M. Arunkumar and S. Prabu, "Implementation of Encrypted Image Compression using Resolution Progressive Compression Scheme," *International Journal of Computer Science and Mobile Computing (IJCSMC)*, vol. 3, no. 6, pp. 585–590, 2014.

[11] A. Razzaque and N. V Thakur, "An Approach to Image Compression with Partial Encryption without sharing the Secret Key," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 12, no. 7, pp. 1–6, 2012.

[12] X. Kang, A. Peng, X. Xu, and X. Cao, "Performing Scalable Lossy Compression on Pixel Encrypted Images," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 32, 2013.

[13] H. K. Aujla and R. Sharma, "Designing an Efficient Image Encryption Then Compression System with Haar and Daubechies Wavelet," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 5, no. 6, pp. 7784–7788, 2014.

[14] Y. M. Kamble and K. B. Manwade, "Secure Data Communication using Image Encryption and Compression," *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, vol. 3, no. 12, pp. 8–11, 2014.

[15] M. Sharma and S. Gandhi, "Compression and Encryption : An Integrated Approach," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 5, pp. 1–7, 2012.

[16] L. Krikor, S. Baba, T. Arif, and Z. Shaaban, "Image Encryption Using DCT and Stream Cipher," *European Journal of Scientific Research*, vol. 32, no. 1, pp. 47–57, 2009.

[17] M. Benabdellah, F. Regragui, and E. H. Bouyakhf, "Hybrid Methods of Image Compression-Encryption," *J. of Commun. & Comput. Eng.*, vol. 1, no. 1, pp. 1–11, 2011.

[18] C. Samson and V. U. K. Sastry, "A Novel Image Encryption Supported by Compression Using Multilevel Wavelet Transform," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 3, no. 9, pp. 178–183, 2012.

[19] C. Samson and V. U. . Sastry, "An RGB Image Encryption Supported by Wavelet- based Lossless Compression," *International Journal of Advanced Computer and Aplications (IJACSA)*, vol. 3, no. 9, pp. 36–41, 2012.

[20] K. Gupta and S. Silakari, "Novel Approach for Fast Compressed Hybrid Color Image Cryptosystem," *Advances in Engineering Software*, vol. 49, no. 1, pp. 29–42, Jul. 2012..

[21] P. Li and K. Lo, "Joint Image Compressio n and Encryption Based on Alternating Transforms with Quality Control," in *2015 Visual Communications and Image Processing (VCIP)*, 2015, pp. 1–4.

[22] Chung-Ping Wu and C.-C. J. Kuo, "Design of integrated multimedia compression and encryption systems," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 828–839, Oct. 2005.

[23] H. Hermassi, R. Rhouma, and S. Belghith, "Joint compression and encryption using chaotically mutated Huffman trees," *Communications in Nonlinear Science and Numerical Simulation (ELSEVIER)*, vol. 15, no. 10, pp. 2987–2999, 2010.

[24] J. Chen, J. Zhou, and K.-W. Wong, "A Modified Chaos-Based Joint Compression and Encryption Scheme," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 58, no. 2, pp. 110–114, Feb. 2011.

[25] P. S. Kishore, N. A. Nagendra, K. P. Reddy, and V. V. S. Murthy, "Smoothing and Optimal Compression of Encrypted Gray Scale Images," *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 3, pp. 23–28, 2012.

[26] A. V. Nair. S, G. K. Sundararaj, and T. S. R. Perumal, "Simultaneous Compression and Encryption using Arithmetic Coding with Randomized Bits," *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 2, no. 2, pp. 38–42, 2012.

[27] Sudesh, A. Kaushik, and S. Kaushik, "A Two Stage Hybrid Model for Image Encryption and Compression to Enhance Security and Efficiency," in *2014 International Conference on Advances in Engineering & Technology Research(ICAETR - 2014)*, 2014, pp. 1–5.

[28] T. Xiang, J. Qu, and D. Xiao, "Joint SPIHT Compression and Selective Encryption," *Applied Soft Computing*, vol. 21, pp. 159–170, Aug. 2014.

[29] S.-C. Ou, H.-Y. Chung, and W.-T. Sung, "Improving the compression and encryption of images using FPGA-based cryptosystems," *Multimedia Tools and Applications*, vol. 28, no. 1, pp. 5–22, Jan. 2006.

[30] B. Chai, J. Vass, X. Zhuang, and C. Science, "Significance-Linked Connected Component Analysis for Low Bit Rate Image Coding," vol. 8, no. 6, pp. 774–784, 1999.

[31] A. Alfalou, C. Brosseau, N. Abdallah, and M. Jridi, "Assessing The Performance of a Method of Simultaneous Compression and Encryption of Multiple Images and Its Resistance Against Various Attacks," *Optics Express*, vol. 21, no. 7, pp. 10253–10265, 2013.

[32] X.-J. Tong, Z. Wang, M. Zhang, and Y. Liu, "A New Algorithm of The Combination of Image Compression and Encryption Technology Based on Cross Chaotic Map," *Nonlinear Dynamics*, vol. 72, no. 1–2, pp. 229–241, Apr. 2013.

[33] X.-J. Tong, P. Chen, and M. Zhang, "A Joint Image Lossless Compression and Encryption Method Based on Chaotic Map," *Multimedia Tools and Applications*, Jul. 2016. A

[34] L. Zhiqianga, S. Xiaoxin, D. Changbin, and D. Qun, "JPEG Algorithm Analysis and Application in Image Compression Encryption of Digital Chaos," in *2013 Third International Conference on Instrumentation, Measurement, Computer, Communication and Control*. IEEE, 2013, pp. 185–189.

[35] N. Goel, B. Raman, and I. Gupta, "Chaos Based Joint Compression and Encryption Framework for End-to-End Communication Systems," *Advances in Multimedia.*, vol. 2014, pp. 1–10, 2014.

[36] M. Kumar and A. Vaish, "An Efficient Encryption-Then-Compression Technique for Encrypted Images using SVD," *Digital Signal Processing*, vol. 60, pp. 81–89, Jan. 2017.

[37] W. M. Rahmawati, A. Saikhu, and A. E. Kompresi, "Implementasi Algoritma Penggabungan Kompresi dan Enkripsi Citra dengan DCT dan SHA-1," *Jurnal Teknik POMITS*, vol. 2, no. 1, pp. 1–4, 2013.

[38] R. Challa, G. Vijaya Kumari, and P. S. Sruthi, "Proficient LWE-Based Encryption using CAN Compression Algorithm," in *2015 Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG)*. IEEE, 2015, pp. 304–307.

[39] S. Al-Maadeed, A. Al-Ali, and T. Abdalla, "A New Chaos-Based Image-Encryption and Compression Algorithm," *Journal of Electrical and Computer Engineering*, vol. 2012, pp. 1–11, 2012.

[40] C. Wang, J. Ni, and Q. Huang, "A New Encryption Then Compression Algorithm using The Rate Distortion Optimization," *Signal Processing: Image Communication*, vol. 39, pp. 141–150, Nov. 2015.

[41] N. S. Hassan and H. A. Younis, "Approach For Partial Encryption Of Compressed Images," *Journal of Babylon University/Pure and Applied Sciences*, vol. 21, no. 3, pp. 1–10, 2013.

[42] J. Xiaoyong, B. Sen, Z. Guibin, and Y. Bing, "Image encryption and compression based on the generalized knight's tour, discrete cosine transform and chaotic maps," *Multimedia Tools and Applications*, Jul. 2016.

[43] G. Alvarez and S. Li, "Some Basic Cryptographic Requirements for

Chaos Based Cryptosystems," *International Journal of Bifurcation and Chaos*, vol. 16, no. 8, pp. 2129–2151, Aug. 2006.

[44] M. Hamdi, R. Rhouma, and S. Belghith, "A Selective Compression-Encryption of Images Based on SPIHT Coding and Chirikov Standard Map," *Signal Processing*, vol. 131, pp. 514–526, Feb. 2017.

[45] N. Zhou, A. Zhang, F. Zheng, and L. Gong, "Novel Image Compression–Encryption Hybrid Algorithm Based on Key-Controlled Measurement Matrix in Compressive Sensing," *Optics & Laser Technology*, vol. 62, pp. 152–160, Oct. 2014.

[46] R. . Huang, K. H. . H. Rhee, and S. . Uchida, "A Parallel Image Encryption Method Based on Compressive Sensing," *Multimedia Tools and Applications*, vol. 72, no. 1, pp. 71–93, Sep. 2014.

[47] M. Fira, "Applications of Compressed Sensing: Compression and Encryption," in *2015 E-Health and Bioengineering Conference (EHB)*. IEEE, 2015, pp. 1–4.

[48] Y. Zhang, K.-W. Wong, L. Y. Zhang, W. Wen, J. Zhou, and X. He, "Exploiting Random Convolution and Random Subsampling for Image Encryption and Compression," *Signal Processing: Image Communication*, vol. 39, no. 20, pp. 202–211, Nov. 2015.

[49] J. Ahmad, M. A. Khan, S. O. Hwang, and J. S. Khan, "A Compression Sensing and Noise-Tolerant Image Encryption Scheme Based on Chaotic Maps and Orthogonal Matrices," *Neural Computing and Applications*, Jun. 2016.

[50] T. Chen, M. Zhang, J. Wu, C. Yuen, and Y. Tong, "Image Encryption and Compression Based on Kronecker Compressed Sensing and Elementary Cellular Automata Scrambling," *Optics & Laser Technology*, vol. 84, pp. 118–133, Oct. 2016.

[51] J. Deng, S. Zhao, Y. Wang, L. Wang, H. Wang, and H. Sha, "Image Compression-Encryption Scheme Combining 2D Compressive Sensing with Discrete Fractional Random Transform," *M Multimedia Tools and Applications*, May 2016.

[52] N. Zhou, S. Pan, S. Cheng, and Z. Zhou, "Image Compression–Encryption Scheme Based on Hyper-Chaotic System and 2D Compressive Sensing," Optics & Laser Technology., vol. 82, pp. 121–133, Aug. 2016.

Model-based Pedestrian Trajectory Prediction using Environmental Sensor for Mobile Robots Navigation

Haruka Tonoki
School of Science for Open and Environmental Systems
Keio University
Yokohama, Japan

Ayanori Yorozu
Keio Advanced Research Centers
Keio University
Yokohama, Japan

Masaki Takahashi
Department of System Design Engineering
Keio University
Yokohama, Japan

***Abstract*—Safety is the most important to the mobile robots that coexist with human. There are many studies that investigate obstacle detection and collision avoidance by predicting obstacles' trajectories several seconds into the future using mounted sensors such as cameras and laser range finder (LRF) for the safe behavior control of robots. In environments such as crossing roads where blind areas occur because of visual barriers like walls, obstacle detection might be delayed and collisions might be difficult to avoid. Using environmental sensors to detect obstacles is effective in such environments. When crossing roads, there are several passages pedestrian might move and it is difficult to depict going each passage in the same movement model. Therefore, we hypothesize that a more effective way to predict pedestrian movement is by predicting passages pedestrian might move and estimating the trajectories to the passages. We acquire pedestrian trajectory data using an environmental LRF with an extended Kalman filter (EKF) and construct pedestrian movement models using vector auto regressive (VAR) models, which pedestrian state is consisting of the position, speed and direction. Then, we test the validity of the constructed pedestrian movement models using experimental data. We narrow down the selection of a pedestrian movement model by comparing the prediction error for each path between the estimated pedestrian state using an EKF, and the predicted state using each movement model. We predict the trajectory using the selected movement model. Finally, we confirm that an appropriate path model that a pedestrian can actually move through is selected before the crossing area and that only the appropriate model is selected near the crossing area.**

Keywords—Prediction of Human Movement; Service Robots; Vector Auto Regressive Models; Kalman Filter; Collision Avoidance

I. Introduction

Various service robots are expected to coexist with humans in real environments. Examples include guidance, communication, and assistant robots. These robots must approach a service user and avoid other humans according to the situation. Especially, for the safe behavior control of autonomous robots that coexist with humans, there are many studies that investigate obstacle detection and collision avoidance using mounted sensors such as cameras and laser range finder. For safety and collision avoidance, several methods have previously been proposed to allow autonomous robots to avoid local collisions reactively: potential field methods [1, 2], social force methods [3], dynamic window approaches [4-6], and vector field approaches [7, 8]. Furthermore, collision avoidance methods for dynamic obstacles such as pedestrian have been proposed which function by predicting obstacles' trajectories several seconds into the future and making decisions based on these predicted trajectories. At present, trajectory prediction methods are important because of the risk of a collision between obstacles and the robot when the trajectory prediction is not sufficiently accurate. With this in mind, this study focused on predicting the dynamic trajectories of pedestrians.

Several methods for predicting pedestrian trajectories assume that pedestrians move with constant speed [9–13]. This assumption may only be effective for short-term predictions because pedestrian trajectories can also change under the influence of the environment. Therefore, some pedestrian trajectory prediction methods considering pedestrian movement tendencies using pedestrian trajectory data that are observed in advance have been proposed.

Those methods predict the trajectory using the current state (pedestrian position and velocity) or the current and previous states. However, a pedestrian's trajectory changes with each step near crossing areas, for example when crossing roads at a crossing point. It may be more effective to consider the pedestrian's state several steps in the past. In this study, we constructed pedestrian movement models based on vector auto regressive (VAR) models. We approximate a pedestrian's position, speed, and direction of movement and predict their trajectory using their states several steps in the past.

Moreover, obstacles may be detected too late to avoid collisions in environments with blind areas caused by visual barriers like walls. In such environments, pedestrian movement prediction methods using environmental sensors are effective [14]. In this study, we constructed a model and predicted the pedestrian's trajectory using an environmental sensor.

It is thought that pedestrians change their direction step by step near environments where multiple passages cross (e.g., when crossing roads). There are many paths to the destination, far from the crossing area. To realize safe mobile robot navigation in such environments, we must construct each path model and predict the pedestrian trajectory, and also evaluate each predicted trajectory and select the appropriate path model for the pedestrian.

This study proposes methods that predict a pedestrian's trajectory, evaluates each predicted trajectory, and selects the

This study was supported by JSPS KAKENHI Grant Number 15H02890.

pedestrian's approaching path using an environmental sensor. We expect that a robot can more effectively avoid pedestrians using this method than existing methods, because it reduces the number of candidate paths near the crossing area.

In concrete terms, we construct pedestrian movement models as follows. First, we acquire pedestrian trajectory data using an environmental LRF with an extended Kalman filter (EKF). Second, we construct VAR models of degree ranging from 2 to 30 for each path. Third, we compare the prediction accuracy for each degree. Then, we decide the pedestrian movement models' degree and verify the constructed models' accuracy. We narrow down the selection of a pedestrian movement model by comparing the prediction error for each path between the estimated pedestrian state using an EKF, and the predicted state using each movement model. Then, we predict the trajectory using the selected movement model. In this study, we verify the validity of the constructed pedestrian movement models using experimental data. Furthermore, we confirm that an appropriate path model that a pedestrian can actually move through is selected before the crossing area and that only the appropriate model is selected near the crossing area.

II. Related Work

Many existing pedestrian trajectory prediction methods use the current state (e.g., pedestrian position and velocity) or the current and previous step states. Shiomi et al. proposed a method that predicts a pedestrian trajectory using the social force model [15]. Similarly, Ratsamee et al. proposed a method that predicts pedestrian trajectories using social force models, considering pedestrian's body pose, face orientation, and personal space [16]. Tamura et al. proposed a method that predicts pedestrian trajectories by storing state transition data in each 1 m^2 and predicting state transitions using the current pedestrian state and the stored state transition data [17]. Tadokoro et al. proposed a method that predicts pedestrian movement by estimating movement tendencies via trial and error when a pedestrian moves in an environmental cell [18]. Noguchi et al. proposed a method that predicts pedestrian movement paths by modeling pedestrian movement between cells based on a variable length Markov model [19]. Other researchers have proposed methods that build pedestrian models using machine learning. Chung et al. used Markov decision processes [20, 21], and Ziebart et al. used a soft-maximum version of Markov decision processes [22]. Callaghan et al. proposed using a Gaussian process [23] and Ellis et al. used Gaussian process regression [24].

These methods do not consider pedestrians' destinations when predicting their movement. However, several methods have been proposed that estimate destination and predict the trajectory toward that destination. Thompson et al. proposed a method that derives the transfer probability of each destination, estimates the destination using random sample consensus, and then predicts pedestrian movement using the derived transfer probability [25]. Bennewiz et al. proposed a method that estimates the destination using an expectation–maximization algorithm, and predicts the trajectory using hidden Markov models [26]. Foka et al. proposed a method that predicts a pedestrian's position at the next step using the current and previous step based on a polynomial neural network. They estimated the destination using the tangent vector of the obstacle's positions at times $t-1, t$ and the predicted position at time $t+1$ [27, 28]. These methods predict indoor trajectories toward destinations such as the TV and the refrigerator. However, when crossing roads where a blind area occurs because of walls, it is difficult to depict taking a right turn, going straight and then taking a left turn in the same movement model. Therefore, we hypothesize that a more effective way to predict pedestrian movement is by predicting pedestrian destination passages and estimating the paths to the passages.

III. Acquirement of Pedestrian Trajectory Data

We conducted an experiment to acquire pedestrian trajectory data when crossing roads using LRF (UTM-30LX, Hokuyo Automatic Co, Ltd., Japan) at the height of 0.22 m that is pedestrian thigh. Figure 1 shows the experimental environment and pedestrian movement direction. We observed the distance to the obstacles at each 0.050 s using LRF in advance. Then, we acquired the position of pedestrian on each time step and collected 157 trajectory data points in total using an EKF with position as the observation value. Table 1 shows the number of trajectory data points that we acquired. We used 142 data points for constructing models and another 15 data point for verifying the models. The process of acquiring the trajectory data is as follows.

To acquire the position data, we compared current LRF data and environmental LRF data that we acquired in advance without the presence of obstacles. Then, we acquired leg data that was different from the environmental LRF data by more than 0.10 m. The $l(i)$-th observed position data point is

$$\boldsymbol{P}_{l(i)} = \left(x_{l(i)}, y_{l(i)}\right) \tag{1}$$

and we cluster by

$$\left\|\boldsymbol{P}_{l(i+1)} - \boldsymbol{P}_{l(i)}\right\| > a \tag{2}$$

thus a is 0.10 m. Figure 2 shows examples of data from one pair of legs, and the derived position at that time. b is the width of the cluster, and the b of the data from each pair of legs is less than 0.20 m. Figure 2 (a) shows the pattern with the legs apart. In this case, we acquire the data as one pair of legs when the adjacent cluster distance c is less than 1.0 m. Figure 2 (b) shows the pattern with the two legs together. In this case, we acquire the data as one pair of legs when b is more than 0.20 m. The position of the pedestrian is at the center of the pair of legs.

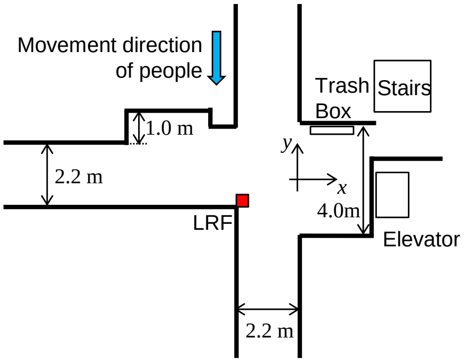
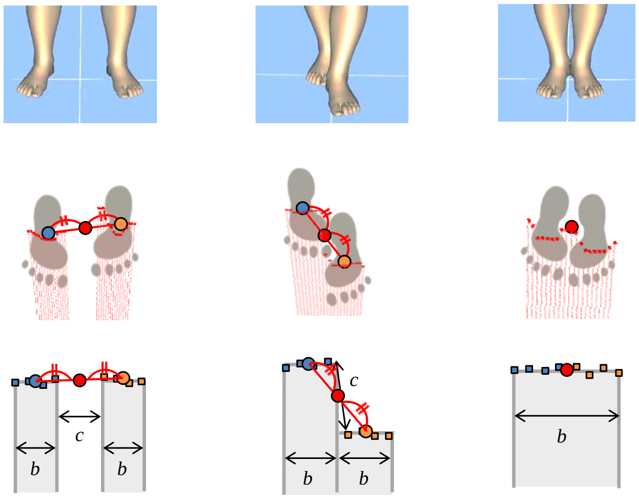


Fig. 1. Experimental environment



(a) Two legs apart (b) Two legs together

Fig. 2. Leg detection

Next, we acquire the trajectory data, which consists of the position data at each time step. Acquired position data includes sensor noise stemming from the LRF accuracy, and the system noise stemming from the position acquisition process. So, we estimate the trajectory data by considering this noise using an EKF [29].

We define the pedestrian state vector $\boldsymbol{x}_k$ at current time step k as:

$$\boldsymbol{x}_k = \begin{bmatrix} x_k & y_k & \theta_k & v_k \end{bmatrix}^T \tag{3}$$

Here, x_k and y_k is position, θ_k is movement direction, and v_k is movement speed. The state equation and observation equation are as follows. :

$$\boldsymbol{x}_k = f\left(\boldsymbol{x}_{k-1}\right) + \boldsymbol{w}_{k-1} \tag{4}$$

$$\boldsymbol{z}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}_k \tag{5}$$

thus the observation value $\boldsymbol{z}_k$ is:

$$\boldsymbol{z}_k = \left[x_k^{LRF} \quad y_k^{LRF} \right]^T \tag{6}$$

and $f\left(\boldsymbol{x}_k\right)$, $\boldsymbol{H}$ is:

$$f\left(\boldsymbol{x}_k\right) = \begin{cases} x_{k-1} + v_{k-1}\Delta t_{k-1}\cos\theta_{k-1} \\ y_{k-1} + v_{k-1}\Delta t_{k-1}\sin\theta_{k-1} \\ \phi_{k-1} \\ \left|v_{k-1}\right| \end{cases} \tag{7}$$

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{8}$$

where $\boldsymbol{w}_k$ is the system noise and $\boldsymbol{v}_k$ is the observation noise.

The prediction and correction steps of the EKF are given by:

Prediction step

$$\boldsymbol{x}_k^- = f\left(\boldsymbol{x}_{k-1}, 0\right) \tag{9}$$

$$\boldsymbol{P}_k^- = \boldsymbol{F}_k \boldsymbol{P}_{k-1} \boldsymbol{F}_k^T + \boldsymbol{Q} \tag{10}$$

where $\boldsymbol{x}_k^-$ is the *a priori* estimation value and $\boldsymbol{P}_k^-$ is the error covariance; and

Correction step

$$\boldsymbol{S}_k = \boldsymbol{H}\boldsymbol{P}_k^- \boldsymbol{H}^T + \boldsymbol{R} \tag{11}$$

$$\boldsymbol{K}_k = \boldsymbol{P}_k^- \boldsymbol{H}^T \boldsymbol{S}_k^{-1} \tag{12}$$

$$\boldsymbol{x}_k = \boldsymbol{x}_k^- + \boldsymbol{K}_k\left(\boldsymbol{z}_k - \boldsymbol{H}\boldsymbol{x}_k^-\right) \tag{13}$$

$$\boldsymbol{P}_k = \boldsymbol{P}_k^- - \boldsymbol{K}_k \boldsymbol{S}_k \boldsymbol{K}_k^T \tag{14}$$

where $\boldsymbol{K}_k$ is the Kalman gain that needs to be calculated, $\boldsymbol{x}_k$ is the *a posteriori* estimate value, and $\boldsymbol{P}_k$ is the error covariance. We define $\boldsymbol{F}_k$, $\boldsymbol{Q}$, $\boldsymbol{R}$ as follows:

$$\boldsymbol{F}_k = \left.\frac{\partial f\left(\boldsymbol{x}, 0\right)}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_k} = \begin{bmatrix} 1 & 0 & -v_k\Delta t_k \sin\theta_k & \Delta t_k \cos\theta_k \\ 0 & 1 & v_k\Delta t_k \cos\theta_k & \Delta t_k \sin\theta_k \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \operatorname{sgn}\left(v_k\right) \end{bmatrix} \tag{15}$$

$$\boldsymbol{Q} = \operatorname{cov}\left(\boldsymbol{w}_k\right) = E\left[\boldsymbol{w}_k \boldsymbol{w}_k^T\right] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\phi^2 & 0 \\ 0 & 0 & 0 & \sigma_v^2 \end{bmatrix} \tag{16}$$

$$\boldsymbol{R} = \operatorname{cov}\left(\boldsymbol{v}_k\right) = E\left[\boldsymbol{v}_k \boldsymbol{v}_k^T\right] = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} \tag{17}$$

where σ_ϕ and σ_v are the variance values of the system noise, and σ_x and σ_y are the variance values of the observation noise. Considering the LRF accuracy and the amount of pedestrian movement change at each time step, we define $\boldsymbol{Q}$ and $\boldsymbol{R}$ as follows:

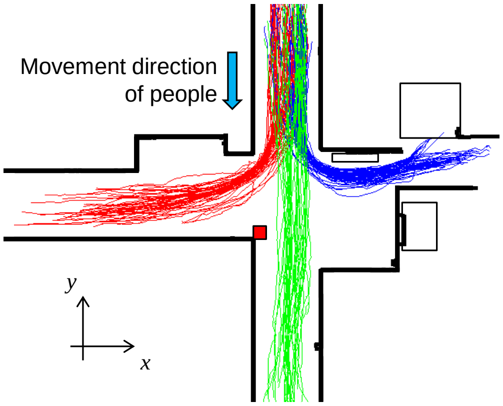


Fig. 3. Estimated human trajectories (red: right trajectory, green: straight trajectory, blue: left trajectory)

$$\boldsymbol{Q} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0076 & 0 \\ 0 & 0 & 0 & 0.000625 \end{bmatrix} \tag{18}$$

$$\boldsymbol{R} = \begin{bmatrix} 0.0025 & 0 \\ 0 & 0.0025 \end{bmatrix} \tag{19}$$

where we acquire the LRF data every 0.050 s.

Figure 3 shows the estimated pedestrian trajectory data using the EKF. In the following, the estimated pedestrian state vector $\boldsymbol{X}_k$ is:

$$\boldsymbol{X}_k = \hat{\boldsymbol{x}}_k = \left[\hat{x}_k \quad \hat{y}_k \quad \hat{\theta}_k \quad \hat{v}_k\right]^T \tag{20}$$

IV. Construction of Pedestrian Movement Models

We construct pedestrian movement models using VAR models. To predict pedestrian trajectory accurately, it is necessary to use high degree models. However, more time is needed for these models to predict trajectories than for lower degree models, because they have to use more time step data. Therefore, it is necessary to construct models in which prediction error is small but degree is low. The construction of the pedestrian movement is as follows:

First, we construct each 2–30 degree VAR model. VAR models (p) enable us to predict the $(k+1)$th step in the state given the state at the k th step:

$$\hat{\boldsymbol{X}}^d_{k+1|k} = \boldsymbol{\beta}^d_0 + \sum_{i=1}^{p} \boldsymbol{\beta}^d_i \boldsymbol{X}_{k-i+1} \tag{21}$$

where d is a direction parameter that can be concretely right (r), straight (s), or left (l).

Second, we derive the coefficient $\boldsymbol{\beta}^d$ using the maximum likelihood method for each degree to compare accuracy. The multidimensional normal distribution of $\boldsymbol{y}$ with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ and degree D is:

$$N(\boldsymbol{y}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})^T\right\} \tag{22}$$

thus $\boldsymbol{X}_k$ is of the 4th degree, the likelihood function and log-likelihood function of $\boldsymbol{X}_k$ that have mean $\hat{\boldsymbol{X}}^d$ and covariance matrix $\boldsymbol{\Sigma}^d$ are:

$$\begin{aligned} & L\left(\hat{\boldsymbol{X}}^d, \boldsymbol{\Sigma}^d\right) \\ & = \prod_{i=1}^{n^d} N\left(\boldsymbol{X}\,|\,\boldsymbol{\mu}^d, \boldsymbol{\Sigma}^d\right) \\ & = \prod_{i=1}^{n^d} \frac{1}{\sqrt{(2\pi)^4\left|\boldsymbol{\Sigma}^d\right|}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{X}-\hat{\boldsymbol{X}}^d\right)\left(\boldsymbol{\Sigma}^d\right)^{-1}\left(\boldsymbol{X}-\hat{\boldsymbol{X}}^d\right)^T\right\} \end{aligned} \tag{23}$$

$$\begin{aligned} & \log L\left(\hat{\boldsymbol{X}}^d, \boldsymbol{\Sigma}^d\right) \\ & = -2n^d \log(2\pi) - \frac{1}{2}n^d \log\left(\det \boldsymbol{\Sigma}^d\right) \\ & \quad -\frac{1}{2}\sum_{i=1}^{n^d}\left(\boldsymbol{X}_i - \hat{\boldsymbol{X}}^d_i\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X}_i - \hat{\boldsymbol{X}}^d_i\right)^T \end{aligned} \tag{24}$$

where n^d is the number of data steps that are used to construct the VAR models, and the number for each direction is:

$$n^r = 2042, n^s = 1621, n^l = 2645 \tag{25}$$

We estimate VAR models' (p) coefficients $\boldsymbol{\beta}^d_0, \boldsymbol{\beta}^d_1, \cdots, \boldsymbol{\beta}^d_p$, which maximize the log-likelihood function, using the maximum likelihood method.

Third, we compare the models of each degree and decide the degree of the pedestrian movement models. The center is approximately 1.1 m from the edge of the passage, considering the general width of the passage is 2.3 m. Therefore, it takes approximately 2.2 s until the end of the avoidance procedure when the speed of the robot is assumed to be 0.50 m/s. We consider that the robot can avoid a pedestrian with enough margins by predicting the pedestrian trajectory up to 3.0 s in the future. Therefore, we predict the pedestrian trajectory up to 3.0 s in the future. Accordingly, we compare position prediction error E^d_k up to 3.0 s in the future and decide the pedestrian movement models' degree.

$$E^d_k = \frac{0.05}{3.0}\sum_{i=1}^{p}\left\|\boldsymbol{P}_{k-i+1} - \hat{\boldsymbol{P}}^d_{k-i+1|k-i}\right\| \tag{26}$$

where $\boldsymbol{P}_k$ and $\hat{\boldsymbol{P}}^d_k$ is the position vector of $\boldsymbol{X}_k$ and $\hat{\boldsymbol{X}}^d_k$.

The state at the $(k+j)$ th step using the data before the k th step is as follows:

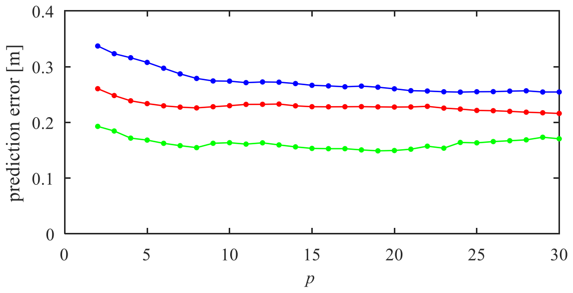


Fig. 4. Prediction error E_k^d until 3.0 s later for each degree (red: right trajectory, green: straight trajectory, blue: left trajectory)

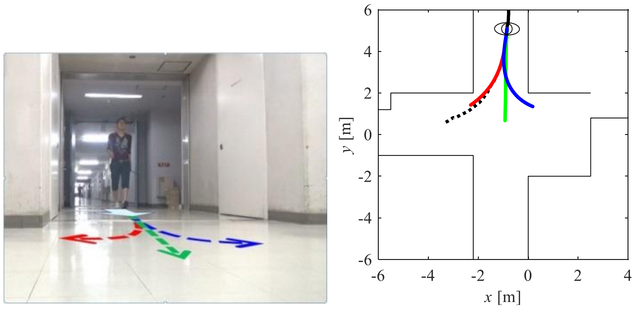
$$
\begin{cases}
\hat{\boldsymbol{X}}^d_{k+j|k} = \boldsymbol{\beta}^d_0 + \sum_{i=1}^{p} \boldsymbol{\beta}^d_i \boldsymbol{X}_{k+j-i} & (j=1) \\
\hat{\boldsymbol{X}}^d_{k+j|k} = \boldsymbol{\beta}^d_0 + \sum_{i=1}^{j-1} \boldsymbol{\beta}^d_i \hat{\boldsymbol{X}}^d_{k+j-i} + \sum_{i=j}^{p} \boldsymbol{\beta}^d_i \boldsymbol{X}_{k+j-i} & (2 \le j \le p) \\
\hat{\boldsymbol{X}}^d_{k+j|k} = \boldsymbol{\beta}^d_0 + \sum_{i=1}^{p} \boldsymbol{\beta}^d_i \hat{\boldsymbol{X}}^d_{k+j-i} & (j > p)
\end{cases}
\tag{27}
$$

Figure 4 shows prediction error E_k^d until 3.0 s in the future for each degree. From Fig. 4, we decide that the pedestrian movement models' degree $p = 8$ because the prediction error decrease after this point is very small.

Next, we verify the appropriateness of the constructed models. The pedestrian trajectory needs to be predicted with about 0.50 m accuracy considering the relative sizes of pedestrians in the environment. From Fig. 4, the constructed models satisfy this prediction accuracy and enable the robot to avoid obstacles safely.

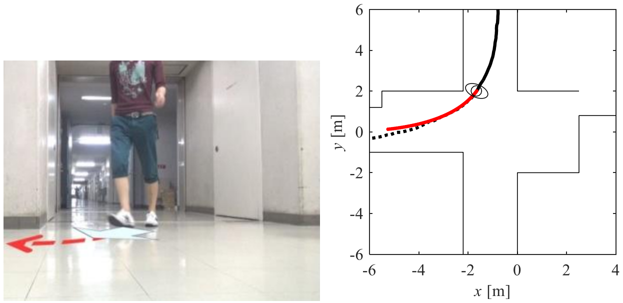
V. Prediction of Pedestrian Trajectory

It is advisable to predict all trajectories that pedestrian might move for safety when crossing roads. It is difficult to predict destination passage pedestrian might move when pedestrian walks far from the crossing area. So, we assume all passages as pedestrian destinations and predict trajectories toward each passage as shown in Fig. 5. However, we can pare down the candidates near the crossing area because pedestrians change their moving direction to the destination. Therefore, we predict toward most likely passage near crossing area as shown in Fig. 6.



(a) Human walking situation (b) Predicted trajectories

Fig. 5. Predicted trajectories when human walks far from the crossing area (black: estimated trajectory, red: predicted right trajectory, green: predicted straight trajectory, blue: predicted left trajectory)



(a) Human walking situation (b) Predicted trajectories

Fig. 6. Predicted trajectories when human walks near the crossing area

TABLE I. Number of Experimental Trajectory Data

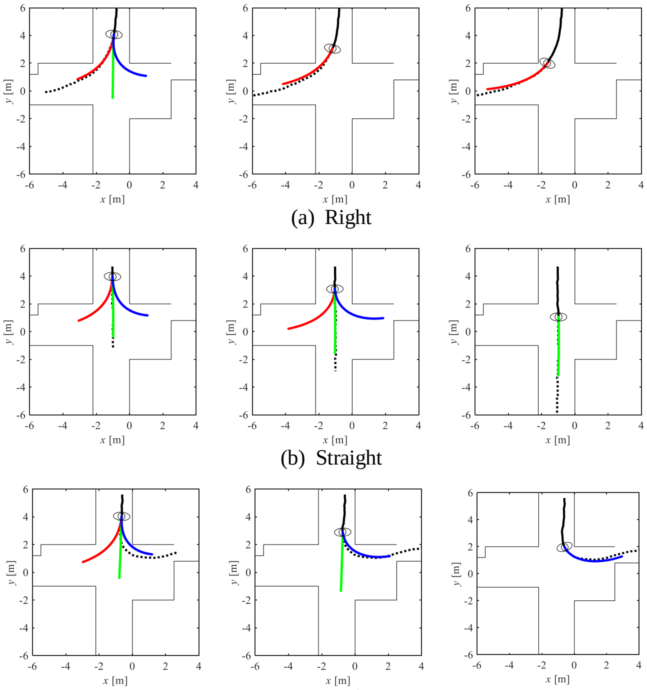
Subject	Model construction	Model verification
Right trajectory data	50	5
Straight trajectory data	35	5
Left trajectory data	57	5
Total	142	15

We predict most likely passages by comparing the average prediction error. First, we calculate average prediction error e_k^d each passages and the minimum average prediction error $e_k^{minimum}$. Second, we select models in which e_k^d is 1.0–1.5 times of $e_k^{minimum}$. e_k^d and $e_k^{minimum}$ are defined as follows:

$$e_k^d = \frac{1}{p}\sum_{i=1}^{p}\left\| \boldsymbol{P}_{k-i+1} - \hat{\boldsymbol{P}}^d_{k-i+1|k-i} \right\| \tag{28}$$

$$e_k^{minimum} = \min\ e_k^d \tag{29}$$

To validate the model against the actual pedestrian trajectory, we used the verification data as shown in Table 1. Figure 7 shows the predicted and actual trajectories. Tables 2 and 3 show the y coordinates at which the appropriate model was chosen.



(a) Right

(b) Straight

(c) Left

Fig. 7. Results of human trajectory prediction

TABLE II. *Y* COORDINATE AT WHICH THE APPROPRIATE MODEL WAS CHOSEN [M]

Subject	Moving direction		
	Right	Straight	Left
Mean	5.75	5.22	5.07
Standard deviation	0.80	0.66	0.87

TABLE III. *Y* COORDINATE AT WHICH ONLY THE APPROPRIATE MODEL WAS CHOSEN [M]

Subject	Moving direction		
	Right	Straight	Left
Mean	1.96	-0.59	1.53
Standard deviation	1.03	0.88	0.34

VI. DISCUSSION

Figure 7 confirms that the number of selected models decreased and that only one appropriate model was selected near the crossing area. The selected models narrowed to only one appropriate model at $y = 2.0$ when turning right (Fig. 7 (a)). However, the narrowing of the selection of models is late when heading straight and turning left (Fig. 7 (b), (c)). Moreover, in Table 2 there is no change in the point at which the appropriate model is selected when going straight or turning right or left, but the point only appreciate model selected is later when heading straight and turning left in Table 3, similar to Fig. 7.

TABLE IV. PREDICTED TRAJECTORY ERROR AT $Y = 1.5$ [M]

Subject	Moving direction		
	Right	Straight	Left
Mean	0.27	0.10	0.30
Standard deviation	0.03	0.01	0.03

TABLE V. PREDICTED TRAJECTORY ERROR AT $Y = 2.0$ [M]

Subject	Moving direction		
	Right	Straight	Left
Mean	0.23	0.20	0.31
Standard deviation	0.03	0.02	0.03

The point at which the selection of models narrows stems from the environment. The environment that we experimented with has a wide road on the right and a narrow road on the left. Moreover, most of the participants whose trajectory was acquired were students who used this environment often, and whose curvatures when turning are thought to be small when turning right and large when turning left. So, there was likely little difference in the position error when heading straight or turning left, because participants tended to begin turning further before the crossing area when turning right than when heading straight or turning left.

Tables 4 and 5 show the means and standard deviations of the prediction error E_k^d at $y = 1.5$ and $y = 2.0$, that is, the mean points where only one appropriate model was selected when turning right and left. We confirm that the constructed models satisfy the prediction accuracy that is necessary for safe obstacle avoidance in an autonomous robot, because pedestrian trajectories need to be predicted with about 0.50 m accuracy considering the size of a pedestrian in the environment.

VII. CONCLUSION

We proposed methods that predict a pedestrian's trajectory, evaluate each predicted trajectory, and select the pedestrian's approaching path using an environmental sensor, for mobile robot navigation. We believe that a robot can avoid a pedestrian with enough margins using the proposed method. Our technique predicts pedestrian trajectories by selecting likely models for environments where several passages cross, and using only one model in environments with only one passage. This method can predict trajectories of several pedestrians if combined with, for example, the potential field or social force methods, and by considering the influence of other pedestrians.

We demonstrated a method to construct pedestrian movement models based on VAR models that consist of pedestrian position, speed and direction for each passage using trajectory data that was acquired in advance by sensors in an environment where a blind area occurs to a mounted sensor on an autonomous robot when crossing roads. We also demonstrated a method of determining the degree of the model, such that degree is kept as low and prediction error as small as possible by comparing prediction error up to 3.0 s in the future. In addition, we validated the accuracy of the constructed models. Furthermore, we showed that we can predict the trajectory in which a pedestrian might move using a movement model that error is lowest between the estimated pedestrian states using an EKF and predicted the upcoming state using each model.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest.

References

[1] Khatib O (1986) Real-time obstacle avoidance for manipulators and mobile robots. Int J Robot Res 5(1): 90-98

[2] Koren Y, Borenstein J (1991) Potential field methods and their inherent limitations for mo-bile robot navigation. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 1398-1404

[3] Helbing D, Molnar P (1995) Social force model for pedestrian dynamics. In: Physical review E 51: 4282

[4] Fox D, Burgard W, Thrun S (1997) The dynamic window approach to collision avoidance. IEEE Robot Autom Mag 4(1): 23-33

[5] Ögren P, Leonard NE (2005) A convergent dynamic window approach to obstacle avoidance. IEEE Trans Robot 21(2): 188-195

[6] Brock O, Khatib O (1999) High-speed navigation using the global dynamic window approach. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 341-346

[7] Borenstein J, Koren Y (1991) The vector field histogram-fast obstacle avoidance for mobile robots. IEEE Trans Robot Autom 7(3): 278-288

[8] Borenstein J, Koren Y (1990) Real-time obstacle avoidance for fast mobile robots in cluttered environments. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 572-577

[9] Ohki T, Nagatani K, Yoshida K (2010) Collision avoidance method for mobile robot considering motion and personal spaces of evacuees. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 1819-1824

[10] Granata C, Bidaud P (2012) A framework for the design of person following behaviors for social mobile robots. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 4652-4659

[11] Tsubouchi T, Arimoto S (1994) Behavior of a mobile robot navigated by an "iterated forecast and planning" scheme in the presence of multiple moving obstacles. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 2470-2475

[12] Belkhouche F (2009) Reactive path planning in a dynamic environment. IEEE Trans Robot 25(4): 902-911

[13] Pacchierotti E, Christensen HI, Jensfelt P (2006) Evaluation of passing distance for social robots. In: Proceedings of IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp 315-320

[14] Batalin MA, Sukhatme GS, Hattig M (2004) Mobile robot navigation using a sensor network. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 636-641

[15] Shiomi M, Zanlungo F, Hayashi K, Kanda T (2014) Towards a socially acceptable collision avoidance for a mobile robot navigating among pedestrians using a pedestrian model. Int J Soc Robot 6(3): 443-455

[16] Ratsamee P, Mae Y, Ohara K, Takubo T, Arai T (2013) Human–robot collision avoidance using a modified social force model with body pose and face orientation. Int J Soc Humanoid Robot 10(1): 1350008

[17] Tamura Y, Hamasaki S, Yamashita A, Asama H (2013) Collision avoidance of mobile robot based on prediction of human movement according to environments. Transactions of the Japan Society of Mechanical Engineers 79(799): 617-628 (in Japanese)

[18] Tadokoro S, Hayashi M, Manabe Y, Nakami Y, Takamori T (1995) Motion planner of mo-bile robots which avoid moving human obstacles on the basis of stochastic prediction. In: Proceedings of IEEE International Conference on Intelligent Systems for the 21st Century, pp 3286-3291

[19] Noguchi H, Yamada T, Mori T, Sato T (2012) Mobile robot path planning using human prediction model based on massive trajectories. In: Proceedings of IEEE International Conference on Networked Sensing Systems (INSS), pp 1-7

[20] Chung SY, Huang HP (2010) A mobile robot that understands pedestrian spatial behaviors. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 5861-5866

[21] Chung SY, Huang HP (2012) Incremental learning of human social behaviors with feature-based spatial effects. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 2417-2422

[22] Ziebart BD, Ratliff N, Gallagher G, Mertz C, Peterson K, Bagnell JA, Hebert M, Dey AK, Srinivasa S (2009) Planning-based prediction for pedestrians. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 3931-3936

[23] Callaghan ST, Singh SP, Alempijevic A, Ramos FT (2011) Learning navigational maps by observing human motion patterns. In: Proceedings of IEEE International Conference on Robotics and Automation, pp 4333-4340

[24] Ellis D, Sommerlade E, Reid I (2009) Modelling pedestrian trajectory patterns with gaussian processes. In: Proceedings of 12th IEEE International Conference on Computer Vision Workshops, pp 1229-1234

[25] Thompson S, Horiuchi T, Kagami S (2009) A probabilistic model of human motion and navigation intent for mobile robot path planning. In: Proceedings of IEEE International Conference on Autonomous Robots and Agents (ICARA), pp 663-668

[26] Bennewitz M, Burgard W, Cielniak G, Thrun S (2005) Learning motion patterns of people for compliant robot motion. Int J Robot Res 24(1): 31-48

[27] Foka AF, Trahanias PE (2002) Predictive autonomous robot navigation. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 490-495

[28] Foka AF, Trahanias PE (2010) Probabilistic autonomous robot navigation in dynamic environments with human motion prediction. Int J Soc Robot 2(1): 79-94

[29] Bellotto N, Hu H (2010) Computationally efficient solutions for tracking people with a mo-bile robot: an experimental evaluation of Bayesian filters. Autonomous Robots 28(4): 425-438

Efficient Model for Distributed Computing based on Smart Embedded Agent

Hassna Bensag, Mohamed Youssfi, Omar Bouattane
Laboratory SSDIA, ENSET
Hassan II University of Casablanca
Mohammedia 28999, Morocco

Abstract—**Technological advances of embedded computing exposed humans to an increasing intrusion of computing in their day-to-day life (e.g. smart devices). Cooperation, autonomy, and mobility made the agent a promising mechanism for embedded devices. The work aims to present a new model of an embedded agent designed to be implemented in smart devices in order to achieve parallel tasks in a distribute environment. To validate the proposed model, a case study was developed for medical image segmentation using Cardiac Magnetic Resonance Image (MRI). In the first part of this paper, we focus on implementing the parallel algorithm of classification using C-means method in embedded systems. We propose then a new concept of distributed classification using multi-agent systems based on JADE and Raspberry PI 2 devices.**

Keywords—***Distributed computing; parallel computing; Multi Agent System; Embedded computing; Raspberry PI 2***

I. Introduction

Technological advances had imposed a growing intrusion of data processing tools as smart devices, giving us the opportunity to grow towards a continuous mobility.

Ambient intelligence does not merely adapt the technology to the human need, but also to the science demands by providing advanced embedded devices with high-level computing power. The low cost of some smart devices like raspberry Pi, and Arduino made them a fertile platform for high performance computing (HPC), an area that was previously very limited due to the cost and the complexity of HPC cluster. Today, thanks to smart devices advanced features, building a cluster to explore parallel computing has become even more cheaper and easier [19]. To fully exploit the cluster resources potential, strong jobs are partitioned into several tasks; these sub tasks are then distributed to multiple smart devices aiming to reduce the cost of communication, latency and execution time. Therefore, introducing some cooperating and social reasoning capabilities to these intelligent devices is necerray.

An intelligent agent is "a computer system, situated in some environment and capable of flexible and autonomous action in order to meet its design objectives" [15,18]. Multi-agent systems are based on the approach: compute corporately and autonomously. Even though multi-agent approach seems appropriate for raspberry devices, we must solve some agent effective implementation issues [20,16,18,17,12]

The multi-agent systems are used in many domains such as economy simulations, renewable energy, computer science and healthcare domain where image segmentation poses several issues [3,4,8,9,10 and 11]. In fact, when the image contains a large amount of data, the segmentation process takes a long time [5, 6].

In this article, we focus on the design of intelligent agents embedded in Raspberry Pi device. Also, the implementation of a parallel and distributed environment consisting of a middleware able to manage a set of embedded mobile agents and to provide a mechanism for load balancing and reducing communication cost. The goal is to overcome the distributed computing challenges and ensure a high-performance computing [13]. This paper aims to propose a new method for c-means classification applied to a cardiac MRI. The latter will be segmented on a parallel-distributed platform based on agents, which are embedded in Raspberry pi devices.

The second section of this paper consists in a review of all methods and tools used in the proposed system. The third section gives an overview of the distributed computational model. The proposed architecture is evaluated in section four with a case study using the distributed c-means algorithm. Section five presents the experiment results. Finally, conclusion and future work.

II. Background

This section details selected methods, approaches and tools used in multi-agent system distributed in embedded devices.

A. Multi Agent System (MAS)

An agent is an encapsulated computer system, situated in an environment, and capable of performing flexible and autonomous action in order to meet its design objectives [9]. The main common agent's characteristics are autonomy, reactivity, proactivity, intelligence, adaptability, collaboration and mobility. Mobile agents have the additional ability to move from one machine to another [2,5,9,10 and 14]. Multi-agent system is used in different areas, offering strong models for complex and dynamic environments representation. MAS can also be used to simulate the behavior of complex computer systems, this simulation models can help designers and developers of complex computational systems. So, the multi-agent based simulation provides a good set of tools to manage complex systems for online resource allocation environments.

B. JADE Agent platform

JADE platform is distributed by Telecom Italia the copyright holder, in open source under the terms and conditions of the LGPL (Lesser General Public License Version 2) license [9]. JADE (Java Agent Development) is the most popular open source framework for the development of multi-agent systems, it is a framework fully programmed in JAVA language. It is a FIPA (Foundation for intelligent Physical Agents) compliant agent platform, composed of multiple containers which host and execute agents. The main goal of JADE platform is especially simplifying development while ensuring standard compliance through a comprehensive set of systems for agents and services [1]. Launching JADE platform triggers at least one container called Main Container, if there is other agents, they are registered with the main container [7] (Figure. 1).

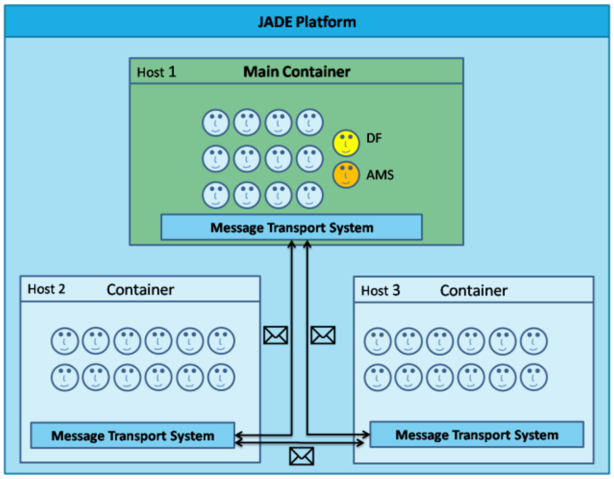


Fig. 1. JADE Agent platform

C. Raspberry PI

Raspberry PI card provides high speed, better accuracy, good flexibility and low cost solution for the development of embedded system equipped by ARM. Using this last board as development platform speed up the process of development. Raspberry pi Model B (as shown in Figure.2) is currently the most popular ARM board. Raspberry PI has a Broadcom BCM2835 system on a chip SoC, which includes an ARM1176JZF-S 700 MHz processor, VideoCore IV GPU, and is shipped with 512MB of RAM .It does not include a built-in hard disk, it uses instead an SD card for booting and long-term storage. It comes with two USB ports, RJ45 Ethernet port, HDMI port and RCA output on board.

Fig. 2. Raspberry PI 2 Model B

D. Distributed C-Means Algorithm

The C-means classification method as defined in [6], is an algorithm for image segmentation consisting of a partitioned groups of set S of n attribute vectors into c classes (clustersC_i , i= 1,…, c), generally based on different criteria segmentation : gray levels, texture or shapes. The main goal of the algorithm is to find the class centers in order to minimize the cost function by using:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \quad \sum_{k \in C_i} d(x_k, C_i)$$

Where:

$C_i$is the center of the i^{th}class

$d(x_k, C_i)$is the distance between i^{th} center C_i and the k^{th} data of the set S

We use the Euclidean distance to define the objective function as follows:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \quad \sum_{k, x_k \in C_i} \| x_k - C_i \|_2$$

The partitioned groups can be defined by a binary membership matrix U(c, n), where each element u_{ij} is formulated by:

$$\begin{cases} 1 \; if \; \| x_k - C_i \|^2 \le \| x_k - C_k \|^2, \forall k \ne i \\ \quad 0 \; otherwise \end{cases}$$

(i=1 tp c, j=1 to n; n is the total number of points in S).

Since a data must belong to only one class, the membership matrix U has two properties which are given in the following equations:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \dots, n$$

$$\sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} = n$$

The value C_i of each class center is computed by the average of all its attribute vectors:

$$C_i = \frac{1}{|C_i|} \sum_{k, x_k \in C_i} x_k$$

$|C_i|$ is the size or the cardinal of C_i.

The C-means classification is achieved using the following algorithm stages as illustrated in figure. 3:

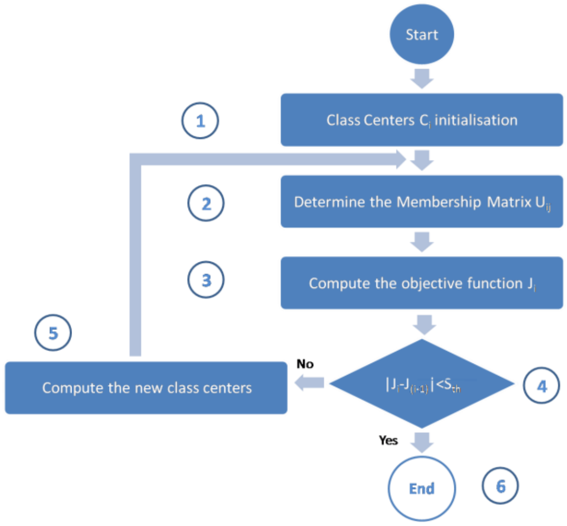

Start
1
Class Centers C_i initialisation
2
Determine the Membership Matrix U_{ij}
3
Compute the objective function J_i
5
Compute the new class centers
No
$|J_i-J_{(i-1)}|<S_{th}$
4
Yes
End
6

Fig. 3. Standard classification c-means process

III. Distributed Computational Architecture

In this section, we present distributed computational architecture which uses agent potential to create a cooperative multi agent platform. The system environment is described before detailing its main components.

A. Computing environment model

The proposed system aims to distribute tasks on a grid of embedded devices: raspberry pi. To perform tasks distribution in the described model (Figure 5), we must achieve the following three steps:

- **Initialization:** the launching of the platform is associated with the creation and initialization of the agent task dispatcher (ATD) using the task prepared by the application.
- **Joining the grid:** once the raspberry pi joins the network it is considered as an available resource and registers itself with the agent task dispatcher to start solving distributed tasks.
- **Task execution:** a task needs a local agent (agent local worker - ALW) and an available embedded remote agent (embedded remote agent - ERA) [15]. Once there are available ERAs, selected from the ATD repository, the execution process starts.

Beginning with task initialization, the process is then followed by task remote execution and ends up with task finalization.

B. Main component description

The proposed platform is a distributed and parallel architecture based on JADE middleware. As shown in figure 4, we distinguish three main components:

- **The main container:** The platform contains one active main container and all other containers joining the platform have to register with it. The main container hosts two special JADE agents: the AMS or agent management system, it keeps the repository of all intelligent agents of the platform. And the DF or directory facilitator; the yellow page service in which agent can register or find the available service in the platform. The main container is also a container where the main agent ATD is deployed.
- **The local container**: is created from the node responsible for the task distribution process initiation in the platform. It is where the ALW agent is hosted.
- **The remote Container**: this container receives groups of the ERAs so that each one can execute its task in parallel.

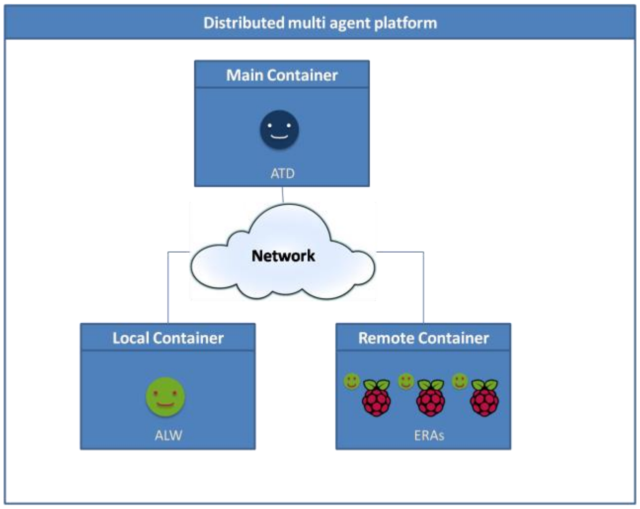


Fig. 4. Distributed multi agent platform

IV. Distributed C-Means Application

Detailed image segmentation according to distributed c-means algorithm is presented in this section. We highlight the segmentation process following the distributed computational model. To handle the raspberry pi characteristics heterogeneity a middleware is required in order to guarantee load balancing.

A. C-means segmentation process in the proposed approach

To prove the reliability of the proposed architecture, we take the c-means image segmentation process as a case study. The proposed architecture uses a C-means algorithm as a distributed program. In order to perform the distributed c-means classification implemented on a grid consisting of embedded devices, we should follow these steps:

- **Task preparation**: we should define the data and processing to be performed for the image c-means segmentation. In this application the user have to choose the source folder containing the image to classify. The classification treatments are defined in a java class in the application.
- **Task initialization**: Once the task is prepared, it is automatically added to the queue of ATD agent before being sent to ALW agent. This latter splits the image based on number (n) of available embedded devices (raspberry pi) provided by the ATD. Each elementary image is then distributed to a specific embedded remote agent (ERA).

- **Task execution**: when the data segmentation is executing, the ALW sends both elementary image and the initial class centers to ERAs. Each ERA determines the local class centers in order to compute the member ship matrix. Afterwards, it replies to the ALW agent message by sending the calculation results gathered from each ERA agent. ALW agent determines then the global class centers and computes the objective function J. The aim is to calculate the absolute value of the difference between J values in iteration i and i-1. If the result is bigger than threshold (Sth), the ALW agent sends a new class centers to the ERAs and the whole process is repeated until obtaining an absolute value lower than Sth. Finally the ERAs send the output elementary segmented images to ALW.
- **Task finalization**: in this step, the LAW assembles the elementary segmented images in order to display the c output segmented images for the classification where c corresponds to the class number.

B. Distributed middleware mechanisms for c-means algorithm

The proposed load distribution middleware aims to develop a multi-agent system to distribute tasks on a grid of embedded devices using intelligent agents. These agents are embedded in heterogeneous nodes of the platform and can dynamically execute the task they receive. We distinguish 3 different types of intelligent agents in this architecture:

- **Agent Task Dispatcher (ATD)**: ATD is the key element of the platform, which has several related functions. First, the ATD overviews the containers and registers agents. It must keep an up-to-date Active Agent Repository of all available intelligent agents. Second, the ATD agent is responsible to distribute tasks among ERAs able to solve tasks. The ATD keeps track of all partial tasks in its Task Allocation tables. Each partial task is marked as “unassigned”, “assigned” or “completed.” Finally, after completing their assigned sub-tasks, ERAs return the partial results to the ALW. After the results have been assembled by the ALW, the ATD deletes the corresponding partial tasks from its Task Allocation tables.
- **Agent Local Worker (ALW):** Any agent on the grid can act as an ALW of a distributed task. If an agent has a task that it cannot solve by itself, it can become an ALW and announce the task to other agents on the grid via the ATD. If there are other available intelligent agents in the grid capable of solving such a distributed task, the task execution process starts, and the ALW takes it over. This agent is the one responsible for achieving both initialization and finalization process.
- **Embedded Remote Agent (ERA):** Any intelligent agent that does not serve as an ALW of a distributed task at a given moment and has registered itself with the ATD is considered as ERA. Once created, these agents move to remote containers where they are supposed to execute their tasks.

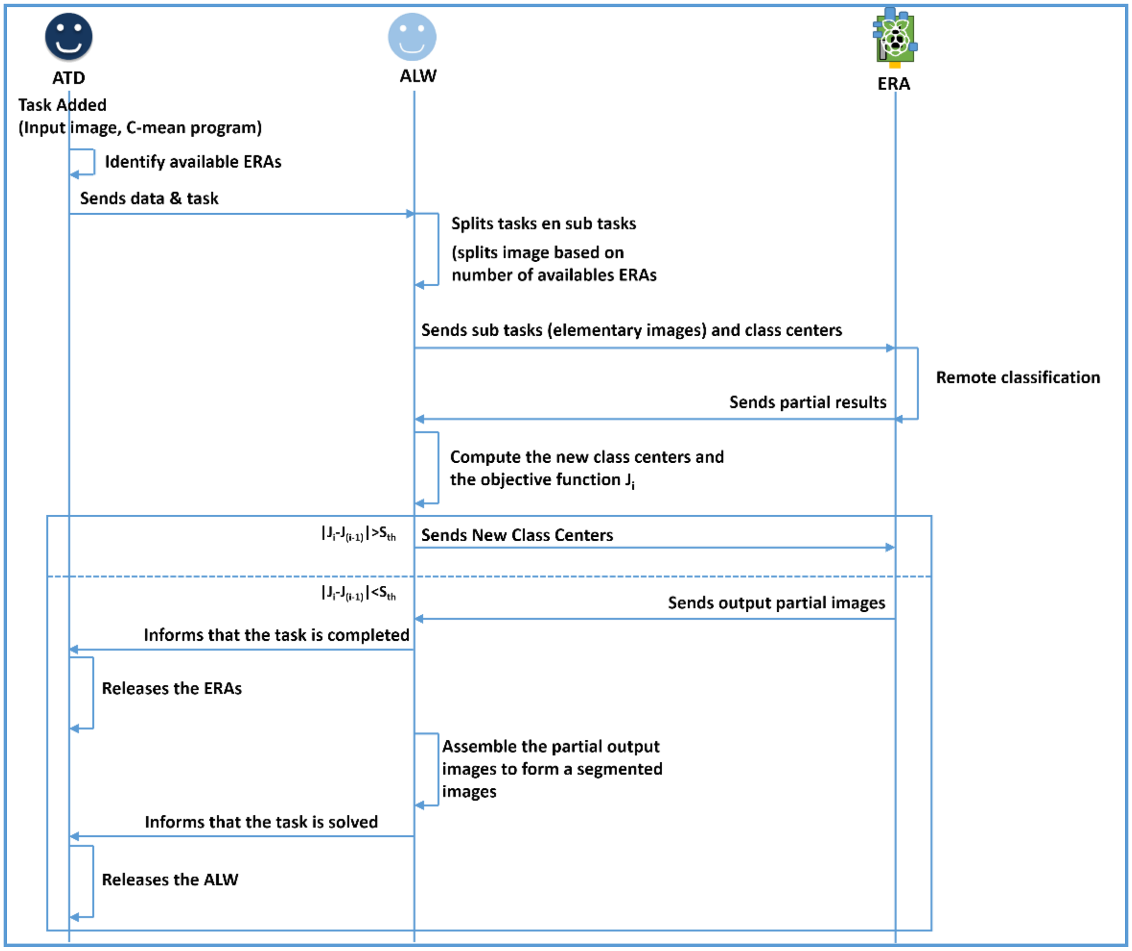

ATD
ALW
ERA
Task Added
(Input image, C-mean program)
Identify available ERAs
Sends data & task
Splits tasks en sub tasks
(splits image based on
number of availables ERAs
Sends sub tasks (elementary images) and class centers
Remote classification
Sends partial results
Compute the new class centers and
the objective function J_i
$|J_i-J_{(i-1)}|>S_{th}$
Sends New Class Centers
$|J_i-J_{(i-1)}|<S_{th}$
Sends output partial images
Informs that the task is completed
Releases the ERAs
Assemble the partial output
images to form a segmented
images
Informs that the task is solved
Releases the ALW

Fig. 5. Distributed Sequence diagram for distributed c-means classification

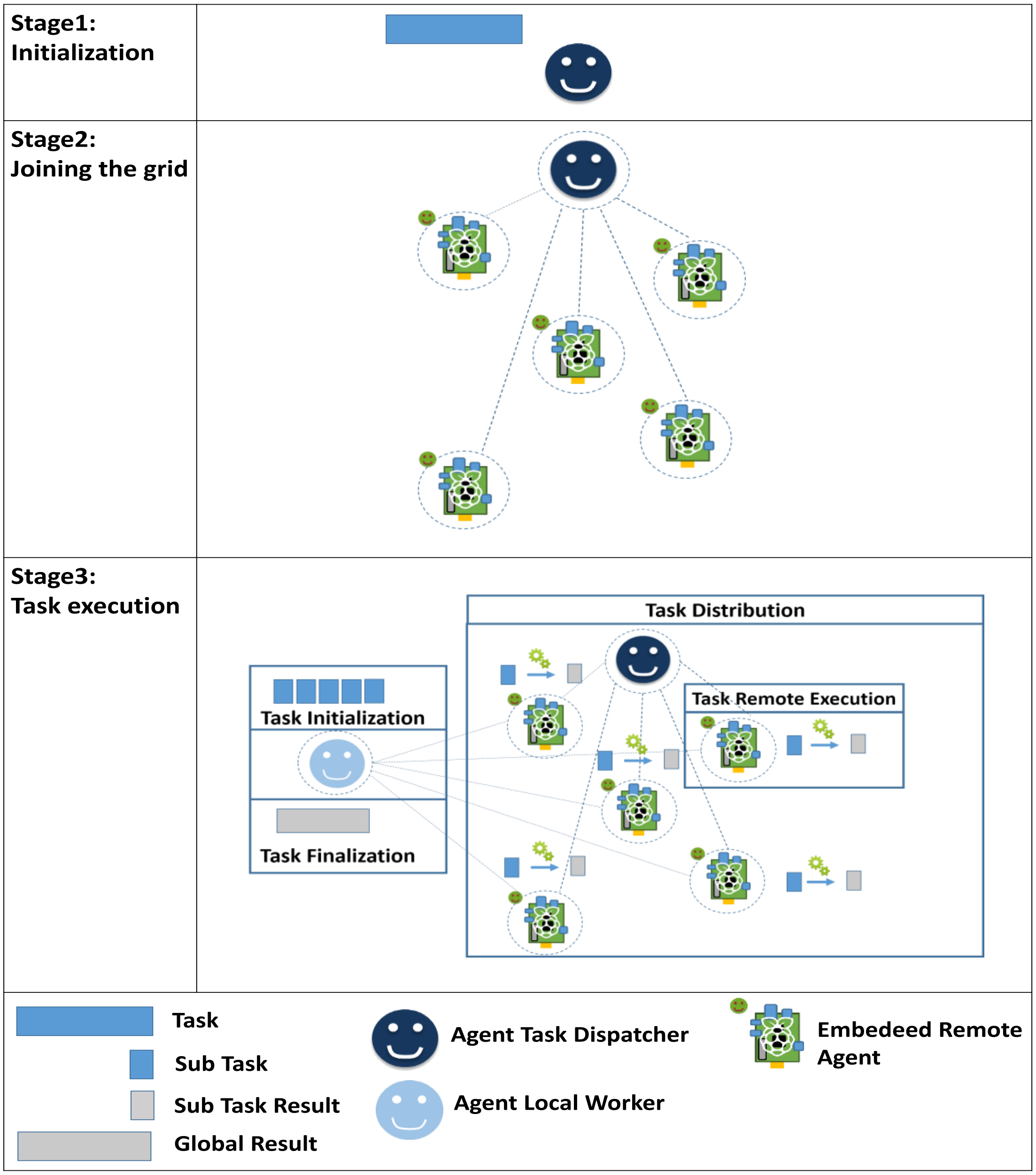


Fig. 6. Distributed model of task execution

V. Simulation and Results

To substantiate the architecture, we implement the c-means program in the platform according to the sequence diagram presented in Figure 6. We have used a cardiac MRI image. The experimental evaluation was done on the test-bed (Figure 7) with two Raspberry PI having the same configuration (Model B) and a third one (Raspberry PI) which host agent. The results are summarized in figure 8: The figure (a) corresponds to a human cardiac MRI, and the figures (b),(c),(d), are the segmented output images where each of them corresponds respectively to class centers (c1=0, c2=127, c3=255).

The first experiment confirms that the behavior of the real distributed systems is coherent with the simulation results, the algorithm converged to the final class centers (c1, c2, c3)=(13.00,99.00,220.00) right after the 8th iteration as shown in Figure 9.

In Figure 10, it's see clearly that from 16 agents the classification time of the two images achieves minimum values of 100 ms. Therefore, we do not need more than 16 devices to obtain this achievement time.

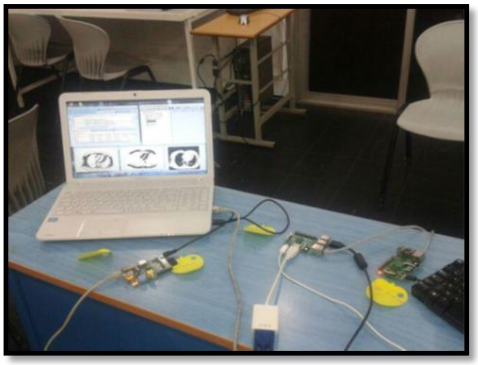


Fig. 7. Testbed

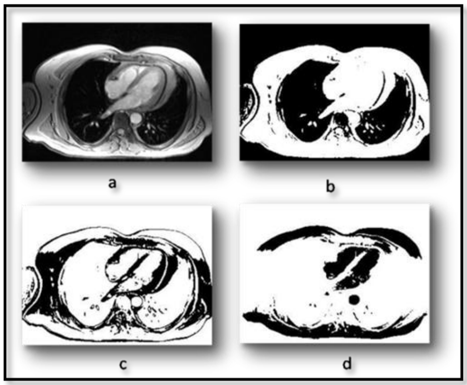


Fig. 8. Segmentation results

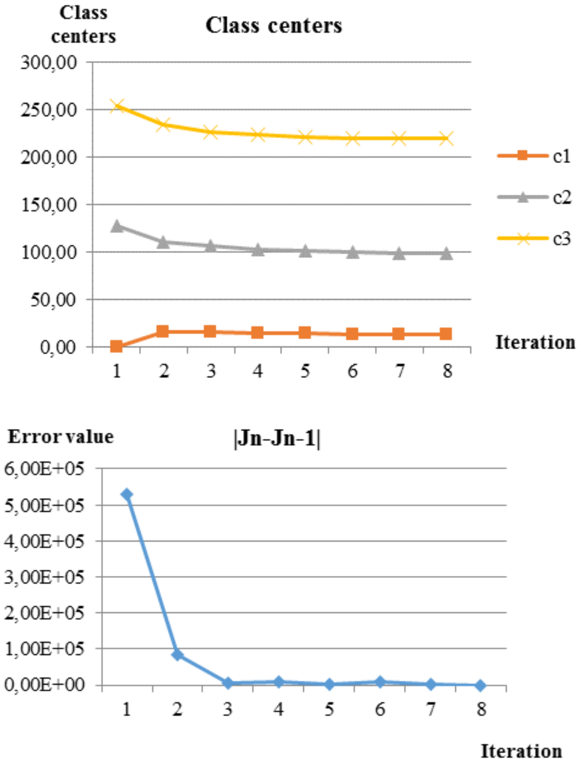


Fig. 9. Class centers (c1, c2, c3)=(0,127,254) And Error of the cost function

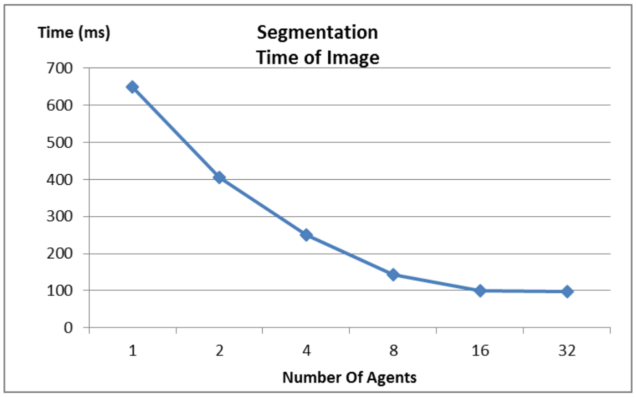


Fig. 10. Many agents by one smart devices

VI. CONCLUSION

The paper presents a distributed computing method with the raspberry PI to fill the gap between the distribution performance and the cluster cost. The proposed architecture can guarantee that, introducing a new model of an ambient agent designed to be implemented in low cost devices, such as raspberry Pi, to achieve parallel tasks in a distributed environment. To experimentally test the efficiency of the proposed system, the method was applied on objective medical image segmented by c-means algorithm, using JADE middleware and embedded agent based on raspberry Pi. The encouraging experiment results proved that the system fills all required features for a performing standard cluster. The proposed architecture opens new horizons towards new and more advanced systems including load balancing and internet of things.

REFERENCES

[1] S.Sotiriadis, N.Bessis, Y.Huang, P.Kuonnen, N.Antonopoulos, A JADE Middleware for Grid inter-cooperated infrastructures, international conference on advanced information networking and applications, 978-7695-4338-3, 2011.

[2] L, Zhang, Q.Wang, X.Shu, A mobile-Agent-Based Moddleware for wirless Sensor Networks Data fusion, International Instrumentation and measurement technology, 978-1-4244-3353-7, May 5-7 Singapore 2009.

[3] L.Chunlin, L.Layuan, A multi-agent model for service-oriented interaction in a mobile grid computing environment, Pervasive and mobile computing 7, 270-284, ELSEVIER, 2011.

[4] M.Youssfi, O.Bouattane, J.Bakkoury, M.O.Bensalah, A new massively parallel and distributed virtual machine model using mobile agents, international conference on multimedia computing and systems, 978-1-4799-3823-0, April 14-16 Marrakech, Morroco 2014.

[5] M. Youssfi, O. Bouattane, and M.O. Bensalah " On the Object Modelling of the Massively Parallel Architecture Computers", Proceedings of the IASTED Inter.Conf. Software engineering, Innsbruck, AUSTRIA, pp 71-78, February 16 - 18, 2010.

[6] O.Bouattane, B. Cherradi, M. Youssfi and M.O. Bensalah "Parallel cmeans algorithm for image segmentation on a reconfigurable mesh computer" ELSEVIER. Parallel computing, 37 pp 230-243, 2011.

[7] F.Bellifemine, A.Poggi, G.Rimassa, Developing Multi-agent systems with JADE, Intelligent Agents VII, pp.89-103, speinger 2001.

[8] M.Higashino, T.Hayakawa, K.Takahashi, T.Kawamura, K.Sugahara, Management of streaming multimedia content using mobile agent technology on pure P2P-based distributed e-Learning system, international conference on advanced information networking and applications, 978-0-7695-4953-8, March 25-28 Barcelona 2013.

[9] F. L. Bellifemine, G. Caire, and D. Greenwood, "Developing Multi-Agent Systems with JADE",Wiley, 2007.

[10] I.Satoh, Mobile Agent Middleware for dependable distributed systems, international conference on informatique technology interfaces, june 27-30, Cavtat, Croatia 2011.

[11] R.Abidar, K.Moummadi, H.Medromi, Mobile device and multi agent systems: an implemented platform of real time data communication and synchronization, international conference on multimedia computing and systems, 978-1-61284-730-6, 7-9 April Ouarzazate, Morocco.

[12] F.Bergenti, G.Caire, D.Gotta, Agenst on the move : JADE for android devices, CEUR workshop proceeding voal-11260, sepy. 25-26 catania, Italy 2014.

[13] Petr Kadera1, Petr Novak1, Vaclav Jirkovsky, Pavel Vrba1, Performance models preventing multi-agent systems from overloading computational resources, Automation, Control and Intelligent Systems, 2(6): 105-111, 2014.

[14] Abhilash Kantamneni, Laura E. Brown, Gordon Parker, Wayne W. Weaver, Survey of multi-agent systems for microgrid control, Engineering Applications of Artificial Intelligence 45, 192–203, ELSEVIER, 2015.

[15] H.Bensag, M.Youssfi, O.Bouattane, Embedded Agent for medical image segmentation, IEEE ICM 2015, 20-23 December Casablanca , Morocco

[16] F.Doctor, H.Hagras, V.Callaghan, "A type-2 fuzzy embedded agent for ubiquitous computing environments", Fuzzy Systems, Proceedings IEEE International Conference , 1105 - 1110 vol.2, July 2004.

[17] T.Leppnen, J.Riekki, M.Liu, E.Harjula, T.Ojala: Mobile agents-based smart objects for the internet of things, Internet of Things Based on Smart Objects, pp. 29–48. Springer, Heidelberg (2014)

[18] H. Hagras, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish and H. Duman, "Creating an ambient-intelligence environment using embedded agents", Intelligent Systems, IEEE, vol. 19, no. 6, pp. 12-20, 2004

[19] C. Ramos, J. C. Augusto and D. Shapiro, "Ambient intelligence - the next step for artificial intelligence", Intelligent Systems, IEEE, 2008

[20] F.Ramparano, O.Boissier, "Smart Devices Embedding Multi-agent Technologies for a Pro-active World", The Ubiquitous Computing Workshop, Bologna, Italy, 16 July 2002.

Decision Framework for Mobile Development Methods

LACHGAR Mohamed
Laboratory of Applied Mathematics and Computer Science
(LAMAI), Faculty of Science and Technology (FSTG),
Cadi Ayyad University
Marrakech, Morocco

ABDALI Abdelmounaïm
Laboratory of Applied Mathematics and Computer Science
(LAMAI), Faculty of Science and Technology (FSTG),
Cadi Ayyad University
Marrakech, Morocco

***Abstract*—Recently, the mobile applications have emerged with the uprising smartphone trend. Now-a-days, a huge number of mobile operating systems require more developments, in order to achieve that, Open source cross-platform mobile frameworks came up, in order to allow importing the same code on various operating systems. In this paper, the focus is made on commonly used mobile development methods, and a process that selects the most suitable solution for a particular need is proposed. Eventually, a new framework that helps to choose the appropriate approach and tool respectively is suggested, according to a convenient survey based on binary questions, in addition to certain criteria.**

***Keywords*—Mobile development approaches; Mobile development tools; Cross-platform mobile; Mobile OS**

I. Introduction

Mobile devices, applications and associated services are being radically reshaped by user’s behavior and corporate organizations as well, either business models, or business strategies and also the way employees work.

Since the release of the first iPhone in 2007, smart mobile devices occupied an important role in the world economy, so we talk more often about digital economy.

Worldwide mobile phones sales reached nearly 478 million units during the third quarter of 2015, so an increase of 3.7 percent compared to the same period in 2014. The figures and the trends presented in the following study confirm these facts [1].

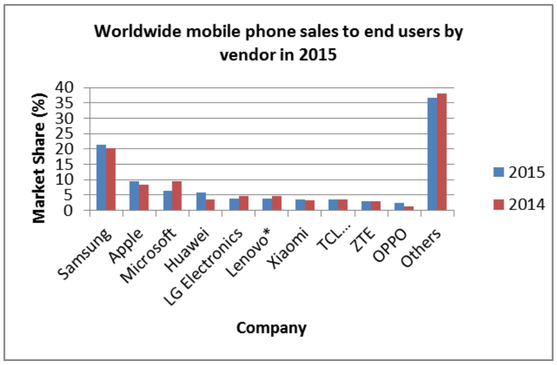


Fig. 1. Worldwide mobile phone sales to end users by vendor in 2015

This evolution is due to the growth the smartphone market, as those consumers abandon more and more “dumb” or “less smart” phones [1]. The following figure shows the evolution of smartphone sales compared to classic mobiles sales.

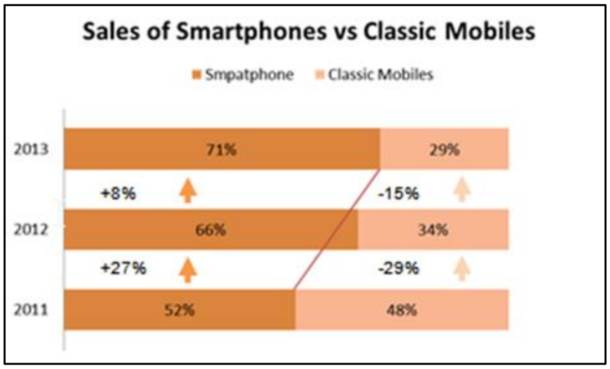


Fig. 2. Sales of smartphones vs classic mobiles [2]

Between 2011 and 2013, the share of smartphone sales increased by 37%. Now-a-days, about 71% of mobiles in markets are smartphones.

The market of tablets and smartphones is dominated by Android [1]. The choice of Android is justified by its constantly innovative technology, open and less expensive compared to iOS.

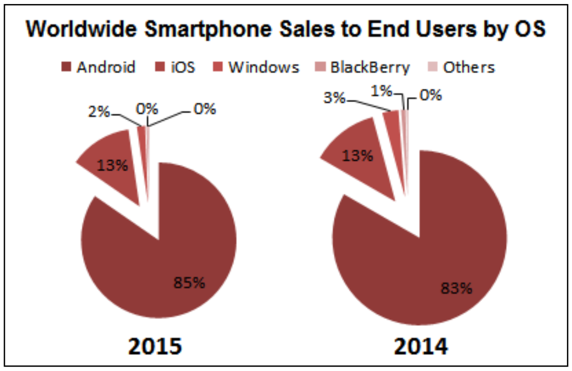


Fig. 3. Worldwide smartphone sales to end users by OS [1]

Each platform indeed requires different development tools. If we want to deploy an application on different platforms, it seems necessary to consume as much time as the sum of the time needed for each application; But there are some solutions to not allow the development of the application once, and then deploy it on other platforms. The aim of this article is, on one hand, to present these solutions and then to make a comparison between them, each one has its advantages and drawbacks; on the other hand, to provide afterwards an ideal approach for deciding which solution should be adopted for a given case.

This paper is structured as follows: the first section presents the mobile development methods, followed by a comparative study of mobile development approaches. The second section shows an ideal approach for deciding which solution should be adopted for a given case. The last section concludes the paper and presents some future works and perspectives.

II. Related Works

Several studies have been carried out on mobile development methods, in which researchers presented the advantages and drawbacks of each approach. In [3] the authors presented a comparative study of multi-platform mobile development tools (PhoneGap, Titanium, Sencha Touch and jQuery Mobile). While the paper [4], has shown the advantages and drawbacks of various methods of mobile development and proposed technologies for each case, based on qualitative properties. However, Charland and Leroux [5] present an in-depth comparison of Native apps and Web apps development. Heitkotter et al. [6] present a comparative study between some cross-platform mobile tools based on several qualitative factors such as licensing costs, look-and-feel, supported platforms, development environments, maintainability and scalability. In this approach, the cross-platform perspective is not taken into account.

Veldhuis [8] present a comparative analysis about the performance of various mobile development tools, based on a simple numerical calculation.

In [9] the authors formulate a method to evaluate and select the best cross-platform development tools for a developer and also evaluate cross platform tools using time, technology, maturity, and cost aspects of mobile apps development. In contrast, this work is focused on the cross-platform development tools and doesn't present a process to assess the appropriate development method to adopt (native, hybrid or web).

In this paper we presented the architecture and features of each method, and an approach that could be adopted to choose an appropriate method and tool is proposed, in order to develop a mobile application.

Our framework focuses on the improvement of decision making in the mobile applications domain, taking into account several qualitative factors such as development rate, documentation, look and feel, popularity, learning curve and graphical tool for GUI. The mentioned framework can be divided into two stages; the first one allows deducing the mobile development method while the second one allows selecting the right tool for each method whose precision exceeds 50%.

III. Mobile Development Methods

The cross platform mobile applications are widely meant to provide mobile apps developers with means for writing once, and deploying everywhere. Currently, the market is full of dizzying array of cross-platform development tools [4].

Several studies on approaches to build cross-platform mobile applications are produced [4], [9], [12], [13]. Conclusively, a classification of these approaches into three categories is made:

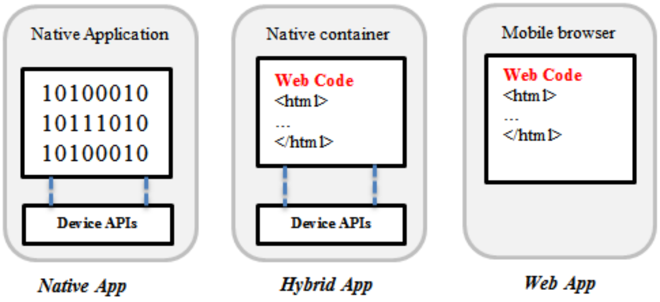


Fig. 4. Mobile development method

These types will be explained in the following sub-sections.

A. Native Approach

The Native applications have the highest performance, native look and feel, has full access to the device capabilities, they use the most updated hardware resources, in order to improve performance. The applications are built in languages that the platform supports, as a consequence it has access to IDEs, which provides the best tools for development, as well as a fast debugging of the project. Android apps can be built in Java on Android Studio, and iOS apps can be built in objective C on XCode, which have all the tools either to debug, or to design the interfaces, and then check the performance using instruments. Yet, the development of the native App needs initial time to learn the languages and tools provided by the platform-specific vendor, then develops the App. Also, the App will run on only one specific-platform [4], [9], [10]. The figure below shows native apps architecture:

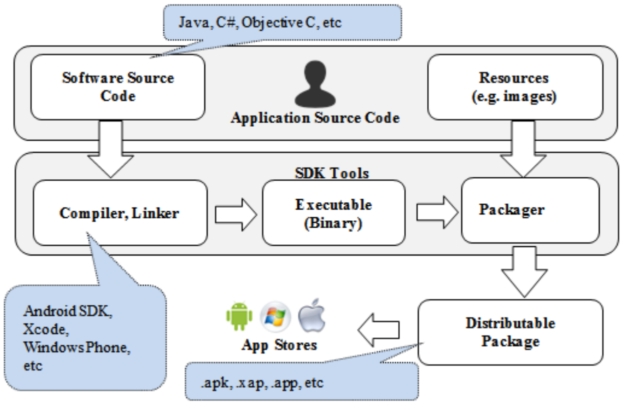


Fig. 5. Native app development

B. Web Approach

The mobile web Apps are developed using standard web technologies—typically HTML5, JavaScript and CSS. These apps are easy to develop, although cannot use device-specific hardware features such as camera or GPS sensor, and the lack look and feel of the native App [11].

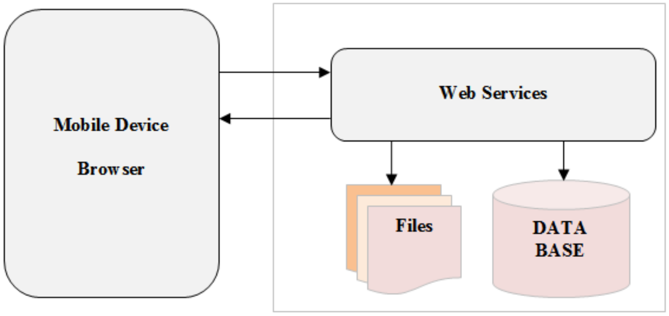


Fig. 6. Logical architecture of a mobile web application [4]

C. Hybrid Approach

The mobile hybrid apps combine between the web App and the native App. This type does not perform as well as the other programs that are based on native languages. Even though they are packaged natively, they are not native applications, they are executed on the platforms web engine, Webkit in case of Android and iOS, which is another layer between the user and the application, and so the performance can't match with the native apps [3], [12].

The below diagram depicts the high level of hybrid mobile application architecture:

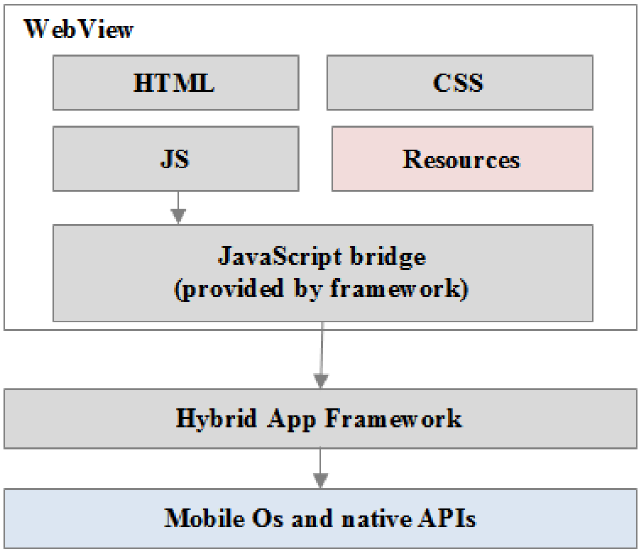


Fig. 7. Logical architecture of a typical hybrid application

D. A comparison of the three approaches

A comparison of the three approaches is structured in the following table.

TABLE I. MOBILE APPS DEVELOPMENT APPROACHES COMPARISON

	Native Approach	**Hybrid Approach**	**Web Approach**
Device Access	Full	Full	Partial
Speed	Very fast	Native speed	Fast
App Development cost	Expensive	Reasonable	Reasonable
AppStore	Yes	Yes	No
Approval Process	Mandatory	Low overhead	None
Quality of UX	Excellent	Not as good as native apps	Very good
Quality of apps	High	Medium to low	Medium
Security	High	Not good	Depends on browser security
Potential users	Limited to a particular mobile platform	Large – as it reaches to users of different platforms	Maximum including smartphones, tablets and other feature phones
Access device-specific features	High	Medium	Low
Development language	Native only	Native and web or web only	Web only
Skills/tools needed for cross-platform apps	Objective-C, Java, C, C++, C#, VB.net	HTML, CSS, JavaScript, Mobile development framework (like PhoneGap)	HTML, CSS, JavaScript

According to this study, native application turned out to be more improved, in terms of performance compared to other mobile application types (i.e., web and hybrid). Native applications are developed using a platform specific API compiled to run on the platform rather than an interpreted language code, such as, JavaScript. But the problem is that these native apps are more expensive to implement, limited to a particular mobile platform, require a collection of knowledge and languages to be realized.

The figure below shows the trend for native to cross platform development cost and time factors.

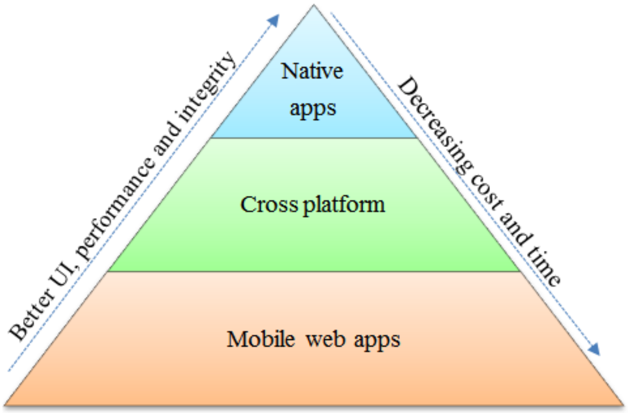


Fig. 8. Native vs. Cross platform development

IV. Decision Framework for Adopting the Appropriate Development Method and Tools

We have shown that the three solutions have advantages and inconveniences. The question now that arises is: which are the approaches that can be adopted to develop a mobile cross-platform application? And what tool can be used to implement the solution?

To answer these questions, a tool to provide answers based on the nature of the application to develop is proposed. The architecture of this tool is presented below (See Figure 9 for more details).

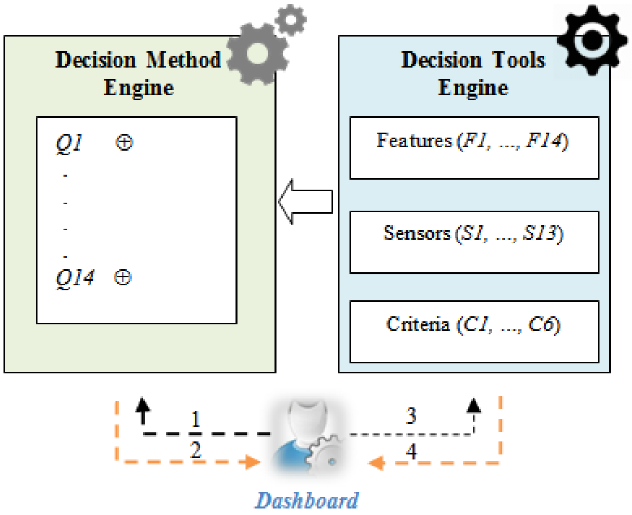


Fig. 9. Framework architecture

In more detail, this architecture consists of four key steps:

1) *The first step :* consists of filling a survey, then sending the answers to the decision engine.

2) *In the second step :* the decision engine analyzes the responses and transmits the appropriate mobile development method to the customer, and also determines the percentage of completion of each method.

3) *In the third step :* according to the received method, the customer must complete a survey, then forward it to the decision tool engine.

4) *In the last step :* the decision tool engine analyzes the responses and sends the right tool to be used in the implementation of the solution to the customer. In the hybrid case the tools are classified according to the features desired in the application to develop.

The next subsections show how each block is implemented.

A. Decision Method Engine

For this, we propose a set of questions in order to single out the correct approach to develop a very specific mobile application.

Q 1 : Should it be published on the main AppStore?
Q 2 : Does it operate in offline mode?
Q 3 : Do you want to sell it?
Q 4 : Is it a simple application?
Q 5 : Will it be frequently used by the user?
Q 6 : Is there an immediate need to deliver the app to the market?
Q 7 : Do you have separate budget for developers in each OS?
Q 8 : Do you need a lot of native features in the Mobile App?
Q 9 : Is app security a high priority?
Q 10 : Should it be very fluid?
Q 11 : Do you want a lot of animations?
Q 12 : Are we building application that needs a lot of algorithmic computation?
Q 13 : Do you want to be always up to date with the latest versions of OS?
Q 14 : Do you want to have the best user experience?

The table below gives the answers to these questions for each mobile development approach (native, hybrid and web).

TABLE II. Mobile Development Methods Decision Framework

	Native	**Hybrid**	**Web**
Q 1	✓	✓	✗
Q 2	✓	✓	✗
Q 3	✓	✓	✗
Q 4	✗	✗	✓
Q 5	✓	✓	✗
Q6	✗	✓	
Q 7	✓	✗	
Q 8	✓	✗	
Q9	✓	✗	
Q 10	✓	✗	
Q 11	✓	✗	
Q12	✓	✗	
Q 13	✓	✗	
Q 14	✓	✗	

In this perspective, we present the selection criterion established in a decision tree represented in the figure 10 below.

B. Decision Method Engine implementation

The decision tree shown in Figure above is used to determine the mobile development approach to be taken within a given situation. The decision method engine will also determine the percentage of completion of each method. To do this, we have adopted the following approach.

We have assigned, a decision factor, to each question, according to its importance. The selected intervals clarify these points:

- 8 : Very important
- 6 : Important

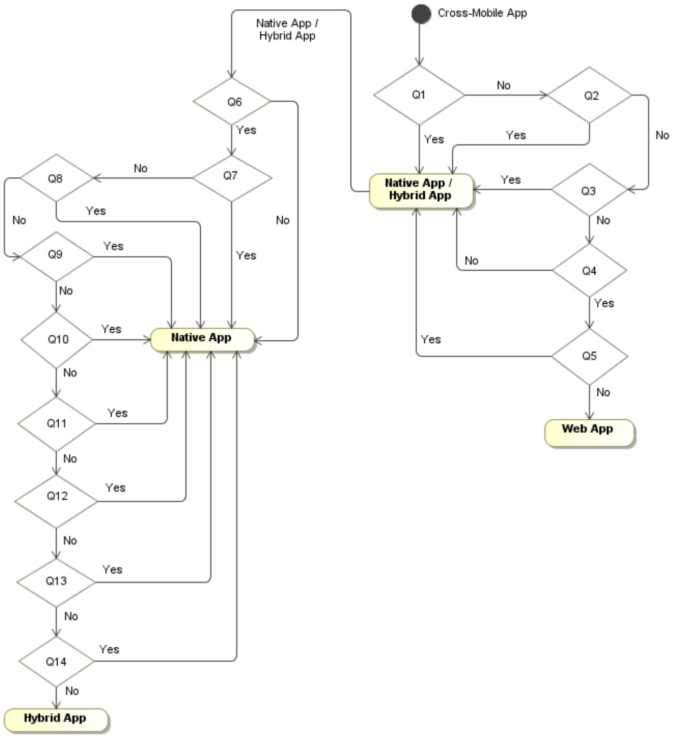


Fig. 10. Decision Tree for adopting the appropriate development method

- 4 : Not so important
- 2 : Not at all important

The chart below illustrates these assigned weights:

TABLE III. FACTORS ATTRIBUTED TO QUESTIONS

Question	Factor
Q1	8
Q2	4
Q3	6
Q4	2
Q5	2
Q6	6
Q7	8
Q8	6
Q9	6
Q10	4
Q11	4
Q12	6
Q13	6
Q14	6

The rationale for choosing various weights of factors is provided below:

- Question 1 is essential, to decide between web approach and the two other approaches. Question 7 is very important for choosing between native approach and the hybrid one, which explains the factor 8 as a decision factor.
- Question 2, actually is less important, especially with the web approach which allows saving data through offline mode according to the HTML 5 innovations.
- Question 3 depends on question 1; a mobile application for sale, must be published in the APPSTORE, thus, we have provided the decision factor 6.
- Question 4 not all important, a simple application can be developed even with all approaches.
- Question 5 not all important, a simple application developed with the web approach, without recourse to the native APIs, can also be used frequently by users.
- Question 6 is important, to decide between native and web approach. If the company has skilled human resources to develop the application within the deadlines set, will be interesting to adopt the native approach, which explains the factor 6 as a decision factor.
- Question 8 can make the difference between native and hybrid approaches. In order to implement an application, an access to several native APIs is required, so it is better to use native approach.
- Question 9 is important, if security is a priority, then it would be better to adopt native approach. Consequently, we assigned 6 as a decision factor for this question.
- Questions 10 and 11 are less important, according to the hybrid approach evolution that supports the implementation of some animations and fluidity depending on the JavaScript Framework evolution, therefore, for both questions, we have assigned the factor 4.
- Whenever the application requires a lot of algorithmic computation, then it is better to use native language for taking advantage of the methods already developed. That explains 6 as a decision factor for a question 12.
- Question 13 is important, if a mobile application has several native features it should benefit from the latest updates of the operating system, which explains the factor 6 as a decision factor.

The feedback is a strong point for this we assigned 6 as a decision factor for a question 14.

An extract of the used class diagram for implementing is shown below:

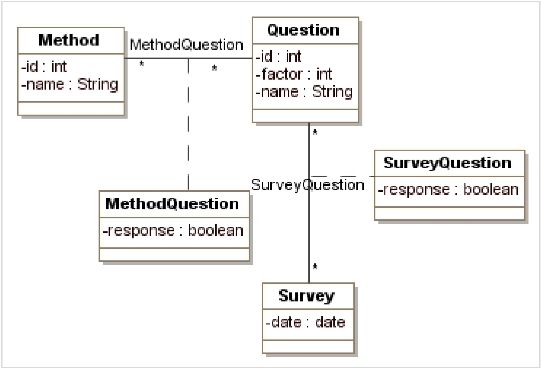


Fig. 11. Extract of class diagram of method engine

The precision is given by the following ratio:

$$\text{Precision (in \%)} = \frac{\sum \text{Factor}(Q_P)}{\sum \text{Factor}(Q_E)} * 100$$

Where:

$\sum \text{Factor}(Q_P)$: is the sum of the factors of the performed questions.

$\sum \text{Factor}(Q_E)$: is the sum of the factors of the expected questions.

With: (Q_P) are the performed questions and (Q_E) are the expected questions (according to TABLE II).

C. Decision Tools Engine

Once the development approach is selected, the next step will be to define the tools to use during the implementation phase. In order to achieve this, we evaluate the needs of the solution to develop towards some sensors and features available in the mobile phone.

The following features and sensors are integrated in many of the major smartphone devices:

TABLE IV. SMARTPHONE DEVICE STANDARD FEATURES

Code	Features	Definition
F1	Contacts	Does the solution supports CRUD functionality to access the contact list?
F2	Geolocation	Does the solution can be capable of using smartphone GPS?
F3	Ad hoc Wi-Fi	Does the solution capable of managing ad hoc Wi-Fi connections?
F4	Storage	Does the solution support CRUD functionality for Local Storage?
F5	SMS	Does the solution have an API to send SMS from the application?
F6	Telephony	Does the solution have an API to make calls from the application?
F7	Bluetooth	Does the solution supply an access to device Bluetooth?
F8	Audio (Recording)	Does the solution allow audio playback in the application?
F9	Audio (Reading)	Does the solution allow audio recording in the application?
F10	Camera (Take photo)	Does the solution allow taking pictures in the application?
F11	Camera (Video Recording)	Does the solution allow the recording of video in the application?
F12	Vibration	Does the solution allow making vibrate the device since the application?
F13	Multi – touch	Does the solution can be capable of capturing the "Gestures" or the "Multi-touch"?
F14	SOAP	Does the solution have an API to manage the SOAP protocol?
F15	Push Notification	Does the solution contain an API to manage "push notifications"?
F16	SQLite	Does the solution integrate the functionality Create, Read, Update, and Delete (CRUD) of SQLite?
F17	Network availability	Does the solution can be capable of checking the availability of the network?
F18	File System	Does the solution provide to access to the device's file system?
F19	Memory management	Does the solution allow to manually managing memory?

TABLE V. SMARTPHONE DEVICE STANDARD SENSORS

Code	Sensors	Definition
S1	Accelerometer	Does the solution allow to access to the accelerometer?
S2	Compass	Does the solution allow to access to the magnetometer or has it an API to create a compass?
S3	Orientation	Does the solution allow detecting the rotation of the device?
S4	Light sensor	Does the solution allow access to the light sensor?
S5	Gravity	Does the solution allow access to the gravity sensor?
S6	Pressure	Does the solution allow access to the pressure sensor?
S7	Gyroscope	Does the solution allow access to the gyroscope sensor?
S8	Proximity	Does the solution allow access to the proximity sensor?
S9	Temperature	Does the solution allow access to the temperature sensor?
S10	Ambient Temperature	Does the solution allow access to the ambient temperature sensor?
S11	Linear Accelerometer	Does the solution allow access to the linear accelerometer sensor?
S12	Magnetic Field	Does the solution allow access to the magnetic field sensor?
S13	Relative Humidity	Does the solution allow access to the relative humidity sensor?

Also, here are some criteria which may be useful in the selection process:

TABLE VI. SELECTION CRITERIA

Code	Criteria
C1	Development rate
C2	Documentation
C3	Look and feel
C4	Popularity
C5	Learning curve
C6	Graphical tool for GUI

The following sub-section describes and evaluates the mobile development tools, towards the different aspects identified above. These tools are classified in three categories: the platform specific development kit, the Cross-platform mobile development and the web tools.

D. Decision Tools Engine implementation

In the case of the web approach, the tools are defined namely HTML5, CSS and JavaScript.

In the native approach, according to the target platforms, tools to be used for each platform can be defined; therefore, the choice will be unique in this case.

In the case of the hybrid approach the decision tools engine must provide a score that will be calculated, based on the number of features and sensors required in the application that are supported by the tool.

Now, to choose the right tool for implementing a mobile software application, we have defined the following scale:

Rating API or Sensor needs:

- 4: Well Supported.
- 2: Supported.
- 0: Not support.

Development rate:

- 3: Very Fast.
- 2: Fast.
- 1: Medium.
- 0: Slow.

Documentation:

- 3 : Very Good
- 2: Good.
- 1: Fair.
- 0: Poor.

Look and Feel:

- 3: Very Good.
- 2: Good.
- 1: Fair.
- 0: Poor.

Popularity:

- 3: Very popular (Very High).
- 2: Popular (High).
- 1: Less popular (Medium).
- 0: Not popular.

Learning curve:

- 3: Very Fast.
- 2: Fast.
- 1: Medium.
- 0: Long.

Graphical tool for GUI:

- 2: Well supported.
- 1: Supported.
- 0: Not supported.

An extract of the used class diagram for implementing is shown in Figure 12 below:

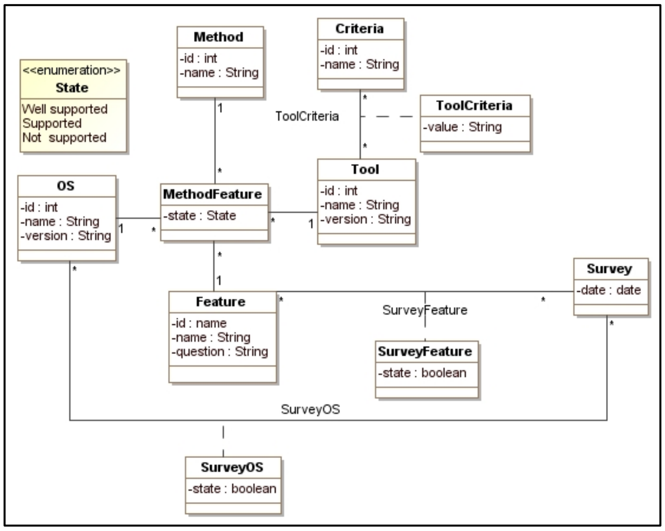


Fig. 12. Extract of class diagram tools engine

V. CASE-STUDY

A. Description:

The aim of this project is to develop a location-based app, this latter allows to locate the position of contacts in the phone book located within a given radius, using a Map, it also provides the ability to communicate with other people connected to the network with the same application, by exchanging text messages and media files (e.g. photo, video), and finally it gives the possibility to take pictures and transmit them via the application to other contacts.

B. Requirements:

- Available on Android and iOS.
- Access to the network.
- Notification Alert and Vibration.
- Access to Camera and video.
- Low costs development.
- Deployable on app stores.
- Access to media.
- Access to Smartphone GPS.
- Access to contacts list.
- Access to telephony.

C. Tools:

- F1 : PhoneGap + jQuery Mobile.
- F2 : PhoneGap + Sencha Touch.

- F3 : PhoneGap + Onsen IU.
- F4 : PhoneGap + Angular UI.
- F5 : PhoneGap + Ionic.
- F6 : Titanium Appcelerator.
- F7 : Xamarin.
- F8 : Flex + Air.

D. *Result:*

- Method decision:

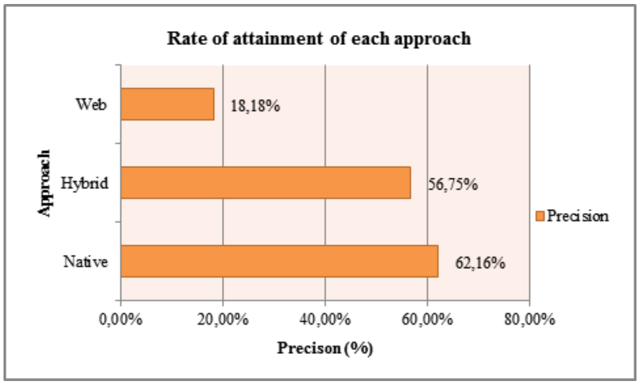


Fig. 13. Rate of attainment of each approach

62.16% of requirements need to adopt the native approach.

56.75% of requirements can be implemented with the hybrid approach.

18.18% of needs can be developed with the web approach.

- Tools decision:

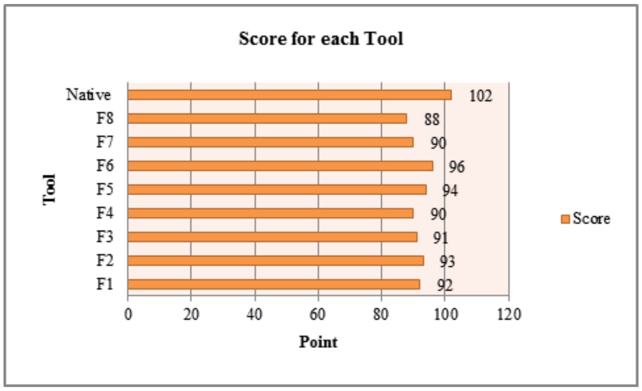


Fig. 14. Score for each tool

For this case-study, the platform specific development kits are among the best, Titanium Appcelerator in the middle followed by PhoneGap with Ionic framework and Sencha Touch, and Flex among the lowest-ranking.

VI. Conclusions and Future Works

This work, presents a framework allowing to select the best technology to use for the development of a specified mobile application in a given context. This framework consists of two main stages, the first one determines the mobile development method (native, hybrid or web) with a completion percentage called precision, based on a set of relevant questions, the second one determines the appropriate tool for the implementation based on a set of relevant criteria.

In an ideal world of technology, without time constraints and money, it would be obviously more interesting to move to a native solution. The result has advantages in terms of ergonomics, performance and integrity.

This study allowed us to understand in which case it is interesting to turn to the web and hybrid solutions. A timely simple and unconstrained performance gain has to be a hybrid or web approach.

Consequently, so as to remedy to native approach's shortcomings, we suggest setting up a solution based on the Model-driven Engineering, allowing developers to generate native applications from the UML diagrams or by using DSL [14], [15].

We are currently working on the development of solutions for reverse engineering, aiming to transform the hybrid code and the web one, into native code. Thus, it will use the native applications advantages and extend them with other native features, which aren't supported now-a-days in the hybrid and web methods.

References

[1] Gartner, “Gartner Says Smartphone Sales Surpassed One Billion Units in 2014”, http://www.gartner.com/newsroom/id/2996817, March 3, 2015 (Accessed on December 3, 2015)

[2] Kerensen Consulting, “Evolution des usages Mobiles, prévision 2015”.

[3] I. Dalmasso, S. Datta, C. Bonnet and N. Nikaein, “Survey, comparison and evaluation of cross platform mobile application development tools”, Proceedings of the 9th International Wireless Communications and Mobile Computing Conference (IWCMC), IEEE Xplore press, Sardinia, pp. 323-328. 2013. DOI: 10.1109/IWCMC.2013.6583580.

[4] R. Raj and S. B. Tolety, “A study on approaches to build cross-platform mobile applications and criteria to select appropriate approach”, India Conference (INDICON), IEEE Xplore press, Kochi, pp. 625-629. 2012. DOI:10.1109/INDCON.2012.6420693.

[5] A. Charland and B. Leroux, “Mobile application development: web vs. native”, Communications of the ACM, vol. 54, no 5, pp. 49-53. 2011.

[6] H. Heitkötter, S. Hanschke and T. A. Majchrzak, “Evaluating cross-platform development approaches for mobile applications”, Web information systems and technologies. Springer Berlin Heidelberg, pp : 120-138. 2012.

[7] L. Delía, N. Galdamez, L. C. Corbalán, P. J. Thomas and P. M. Pesado, “Un análisis comparativo de rendimiento en aplicaciones móviles multiplataforma”, XXI Congreso Argentino de Ciencias de la Computación. 2015.

[8] M. M. O. Veldhuis, “Multi-Target User Interface design and generation using Model-Driven Engineering”, Unpublished dissertation in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Master of Science in Human Media Interaction Faculty of Electrical Engineering, Mathematics and Computer Science University of Twente, the Netherlands, 2013.

[9] P. Smutny, “Mobile development tools and cross-platform solutions”. Proceedings of the 13th International Conference on Carpathian Control (ICCC), IEEE Xplore Press, High Tatras, pp. 653-656, 2013, DOI:10.1109/CarpathianCC.2012.6228727.

[10] S. Xanthopoulos and S. Xinogalos, “A comparative analysis of crossplatform development approaches for mobile applications”, Presented at the Proceedings of the 6th Balkan Conference in Informatics, Thessaloniki, Greece, 2013.

[11] L. Corral , A. Janes and T. Remencius, “Potential Advantages and Disadvantages of Multiplatfonn Development Frameworks-A Vision on Mobile Environments”, Procedia Computer Science, vol. 1 0, pp. 1202-1207, 2012.

[12] M. Palmier, I. Sing and A. Cicchetti, “Comparison of cross-platform mobile development tools”, Proceedings of the 16th International Conference on Intelligence in Next Generation Networks (ICIN), IEEE Xplore Press, Berlin, pp. 179-186, 2012, DOI:10.1109/ICIN.2012.6376023.

[13] N. Serrano, J. Hernantes and G. Gallardo, “Mobile Web Apps”, IEEE Software. pp: 22 – 27, 2013, DOI:10.1109/MS.2013.111.

[14] M. Lachgar and A. Abdali, “Generating Android graphical User Interfaces using an MDA approach”, Proceedings of the Third International Colloquium of Information Science and Technology (CIST), IEEE Xplore Press, Morocco, pp. 80-85, 2014, DOI:10.1109/CIST.2014.7016598.

[15] M. Lachgar and A. Abdali, Abdelmounaïm, “Modeling and generating native code for cross-platform mobile applications using DSL”, Intelligent Automation & Soft Computing, pp. 1-14, 2016.

Prolonging the Network Lifetime of WSN by using the Consumed Power Fairly Protocol

Ahmed Jamal Ahmed
Communication Engineering Department
Universiti Tun Hussein Onn Malaysia,
Parit Raja, Malaysia 86400

Jiwa Abdullah
Communication Engineering Department
Universiti Tun Hussein Onn Malaysia,
Parit Raja, Malaysia 86400

***Abstract*—In wireless sensor networks (WSN), energy saving is always a key concern. Since nodes have limited power, some of them may use specific routes, thus leading to exhaustion of intermediate nodes. These nodes die, which results in routing holes in the network. Consequently, the overall throughput of the network may get reduced. Therefore, in this study, the mathematical proposed model is leaded to optimal route by depending on equal power consumption from whole nodes in the network. Moreover, a new routing protocol model was designed. The protocol is termed as Consumed Power Fairly (CPF). This protocol could achieve high power efficiency by distributing power consumption equally to all nodes in the network. Our proposed model works on finding the route to destination with high power availability after summation of the total power for all nodes from the source to the destination node, thus subtracting the power consumption for particular data required to send. In short, the proposal CPF protocol reduces the number of dead nodes and keeps the connectivity high, which increases prolonging the network lifetime.**

Keywords*—WSN; network topology; energy consumption; graph theory; Consumed Power Fairly

I. INTRODUCTION

Wireless sensor network (WSN) consists typically of multiple autonomous, tiny, low cost and low power sensor nodes. These nodes gather data about their environment and collaborate to forward sensed data to centralized backend units called base stations or sinks for further processing [1]. WSN is known as a kind of wireless ad hoc networks [2]. In sensor networks, energy saving is always a key concern due to various reasons. Some of environments might be very dangerous [3]. Supplying energy through batteries, solar cells energy, other local energy sources are viable options though such challenges are limited to prolong the lifetime of WSN [4]. Therefore, it is important to minimize power requirements across all levels of the protocol stack and minimize the amount of message passing for network control and coordination.

In the field of WSN, recharging is almost impossible due to the small size required for sensors, and large physically distributed networks increase the difficulty of changing batteries [5]. Therefore, researchers and system developers toured to develop the WSN architecture or network layer to minimize energy consumption in order to make the network and overall system application more energy efficient.

The IEEE 802.15.4 standard is specifically meant to support long battery life time [6]. Yet, there are still some precautions to be taken by which a sensor network system application that is based on the standard can run for a longer period of time [7]. Thus, routing is recognized as an essential feature of every network because it is the backbone of a network and it is in charge of forwarding packets among nodes [8]. The routing protocol layer plays an important role in improving the performance of wireless sensor networks for fixed and mobile nodes in the network [9, 10]. Forwarding an amount of data from one node to sink node through intermediate nodes leads to fast death of nodes, especially those nodes which are near to the sink node [3, 11-14]. This will increase the possibility of the whole network death. Therefore, in this study, the demonstrated routing protocol for WSN that is capable of achieving high power efficiency by distributing power consumption equally to all nodes in the network. The sensor node near the sink consumes more power than others, which leads speedy death of nodes and shortening the network lifetime.

As previously stated, in the field of WSN, nodes have limited power, and therefore, some nodes use specific routes, which exhausts the intermediate nodes [15]. Consequently, these nodes die, thus resulting into routing holes in the network that harm the overall network. Furthermore, several problems such as (Shortest path problems, Network flow problems, Matching problems, 2-SAT problem, Traveling Salesman Problem (TSP), and many more) can be formulated and solved by graphs. They use different methods to calculate and select the optimal route from a source to a destination node. To achieve a prolonging lifetime of WSN by fining the optimal path that depends on distributing power consumption on the whole WSN network equally, it is sufficient to achieve energy saving over the whole network. Thus, the study aimed to design a routing protocol for WSN that is able to achieve high power efficiency by distributing power consumption equally on all nodes in the network.

This paper is organized as follows. Section II introduces the mathematical model of energy consumption for our new routing protocol used to find the optimal path. Followed by section III, triangular matrix table to representing a computer network mathematically and graphically. The network topology is represented using a triangular matrix to obtain full topology information. This is followed by Section IV presents the experimental modeling to measure real energy consume to sending and receiving data packets by using CC2420 device (IEEE 802.15.4). As result section V packet energy consumption explain the routing implementation and how to calculate power consumption for each path. Also discusses the method of finding a route inside the WSN. It concludes and shows the simulation and experimental results obtained by using MATLAB.

II. Mathematical Model of Energy Consumption for Routing

In this section, the description of mathematical model used to find the optimal path while transmitting data packets. The sensor node should know the measured value from the environment by reading the sensor value, and then, it should send them to the sink by multi hops and sleep till the next process iteration takes place. Our proposed model depends on finding the route to the destination with high power available after summation of the total power for all nodes from the source to the destination node, thus subtracting power consumption for particular data required to send. Our proposed model depends on network topology.

A graph is a data type structure, which has two components: vertices or nodes and the edges or links that connect them. Graphs can either be undirected or directed. Undirected graphs comprise a group of nodes and a group of links with no direction between a pair of nodes. A group of mobile nodes in a specific domain forms a network that can be represented by a graph. The edges represent the physical wireless connection link among devices, and allow physical transmission between nodes. Topology is defined as a mathematical study of shapes and spaces as represented in (1) where G (N, E) can be the topology graph, N can be the nodes, E stands for the links, and (i, j) refer to the link from Node *i* to Node *j*. Let the information be transmitted by f. The variable vector $X_{ij}^{f}$is then defined as follows:

$$X_{ij}^{f} = \begin{cases} 1 \quad if\ link\ (i,j)\ is\ used\ for\ flow\ f \\ 0\ \ if\ \ link\ (i,j)\ is\ not\ used\ for\ flow\ f \end{cases} \qquad (1)$$

The above equation indicates that the variable vector X_{ij}^{f} is one if the link (i, j) is used to transmit the flow f. Otherwise the vector will be zero if it cannot be used for transmission [16]. As shown in Fig. 1, in order to reach Node 2, only two possible routes exist from Node 5: through the first path ((5, 1) and (1, 2)) and through the second path (5,4), (4,3), and (3, 2)).

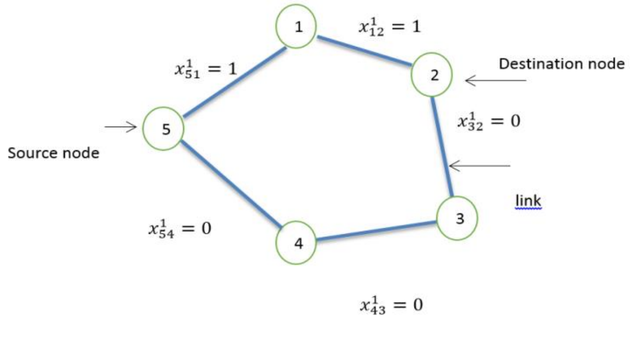


Fig. 1. Graph of routes from Nodes 5 to 2 [17]

Suppose that the X_{ij}^{f} represent the power available in each node. So, to calculate the total summation power available for the first path while X_{51}^{1} = 1, and X_{12}^{1} =1, first, the current power saving (battery) in each node must be read. Assuming the destination node is a sink node that has the infinity power (no battery). So that, it have to sum the power remaining $X_{51}^{1} + X_{12}^{1}$ in Node i=5, and i=1 only ($X_{51}^{1} = power\ remaining$, and $X_{12}^{1} =\ power\ remaining$). In general, (2) shows the total power available to a specific path from the source to destination,

$$T.P = \left\{ \sum_{ij}^{h} \mathrm{X}_{\mathrm{ij}}^{\mathrm{f}}\ (NodePower_i) \right\} \qquad (2)$$

Where T.P denotes the total power available for one path, h represents the number of hops, and h= n-1 where n represents the number of nodes for that particular path.

In assuming that P.N denotes the total power needed (required) to transmit particular data from the source to the sink node, then, power consumption will depend on the receiver and transmitter circuit ($E_{TX,}E_{RX}$). The number of particular transmitted data will also be equal to the number of hops h, denoted as t=h, but for receiving is r=h-1. From the example above, Node 5 transmits only, whereas Node 1 receives and transmits as well. However, to transmit k bits with distance d, the energy consumed is expressed as follows:

$$E_{Tx}(k,d) = E_{Tx\rightarrow elec}(k) +\ E_{Tx\rightarrow amp}(k,d)$$
$$= \begin{cases} kE_{elec} + k\epsilon_{fs}d^2 \quad if\ d < d_0; \\ kE_{elec} + k\epsilon_{mp}d^4 \quad if\ d \geq d_0; \end{cases} \qquad (3)$$

Where d_0 is the distance threshold for swapping amplification models, which can be calculated as $d_0 = \sqrt{\frac{\epsilon fs}{\epsilon mp}}$. Then, the radio will consume power for a k-bit message through the receiving process.

$$E_{Rx}(k) = kE_{elec} \qquad (4)$$

The total power needed for one path is calculated as.

$$P.N = \sum_{t=1}^{t} E_{Tx}(k,d) + \sum_{r=1}^{r} E_{Rx}(k) \qquad (5)$$

If supposed that there are a several of paths to the sink node, then the (2) and (3) should be vectors T.P =[path1, path2,..path$_n$] and P.N=[path1, path2,..path$_n$]. Subtracting T.P - P.N, collecting the remaining power in particular path within the vector as shown in (6).

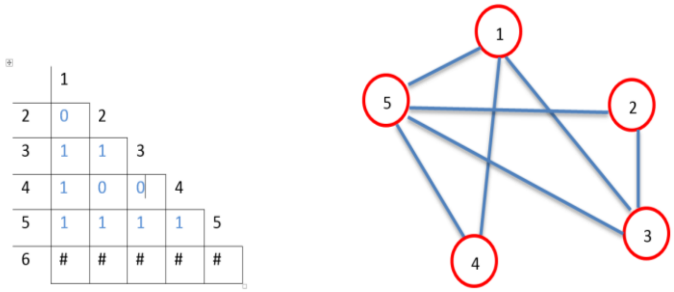
$$P.R = \sum_{p=1}^{p} \left\{ \sum_{ij}^{h} X_{ij}^{f}\ (NodePower_i) - P.N \right\} \qquad (6)$$

Where, P.R represents the residual power for every possible path from the source to the sink node, and p represents the number of possible paths. Hence, P.R represents the vector of elements that give all possible residual power from the source to the sink node. The maximum number represents the optimal path to be selected to achieve a prolong network lifetime. Max

{ P.R: optimal path i }. The sender node must find an optimal route based on network topology and power saving as illustrated in (6). Therefore, the next section explains how to find all possibly available paths based on the network topology.

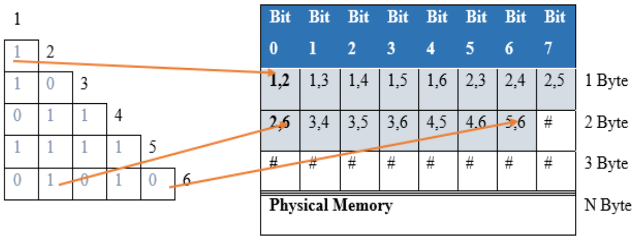
III. Triangular Matrix Table (TMT)

After representing a computer network mathematically and graphically, the network topology is represented using a triangular matrix to obtain full topology information. N is assumed to be the set of nodes and E is the set of links, where n denotes the number of network nodes, that is, n = |N|. Inside the network is a source node s ∈ N and a destination node t ∈ N, and (i, j) ∈ E is the link from Node i to Node j. Using the Triangular Matrix Table (TMT) denotes saving the whole network topology information in a small memory size. First, dimensions of a triangular matrix are equal to the number of nodes, and each node inside the network has a number that represents the diagonal of the lower matrix. At first, the content of the lower matrix must be empty (zero). Each link between two nodes represents one bit inside the lower matrix. Otherwise, the lack of link between the two nodes means that each link will be represented by a zero bit inside the lower matrix depending on the cross of the row node with the column node for nodes with links. Fig. 2 shows the network topology which can be represented mathematically by using digit pairs: (1,2) , (1,3) , (1,4) , (1,5) , (2,3) , (2,4) , (2,5) , (3,4) , (3,5) , (4,5).



	1				
2	0	2			
3	1	1	3		
4	1	0	0	4	
5	1	1	1	1	5
6	#	#	#	#	#

Fig. 2. Representation of Triangular Matrix Table (TMT) and undirected graph



1					
1	2				
1	0	3			
0	1	1	4		
1	1	1	1	5	
0	1	0	1	0	6

Bit 0	Bit 1	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7	
1,2	1,3	1,4	1,5	1,6	2,3	2,4	2,5	1 Byte
2,6	3,4	3,5	3,6	4,5	4,6	5,6	#	2 Byte
#	#	#	#	#	#	#	#	3 Byte
Physical Memory								N Byte

Fig. 3. Representation of triangular matrix table (TMT) in physical memory

To find the routes from the source node to destination, assuming the source node is Node 1 and that the destination node is Node 4. In addition, the TMT is the one shown in Fig. 3. To find the path from Node 1 to Node 4, all the connections with Node 1 have to be read. Therefore, every bit in Column 1 (1, 1, 0, 1, 0) and the corresponding connection related to Nodes (2, 3, 4, 5, and 6) must be checked because "1" bit in the TMT only represents the links. Therefore, the links inside the vector queue must be saved. Consequently, the vector queue becomes (2, 3, 5). The functional check works on each element before adding that element to the vector queue. Moreover, the unique function avoids insertion of any double node number inside the vector queue. Technically, the unique function is a key to ensure a loop-free routing protocol. In addition, the queue mechanism rule is "first in, first out" when saving elements in the vector queue. If the check function does not find the destination node, then, the first element from the vector queues checkouts, which underlies deletion from the vector queue. It also becomes the next step of the search. Here, the vector queue becomes (3, 5, 4, 6). At this point, the check function finds the destination node. Therefore, addition of the node to the queue can be neglected. The routes are 1, 2, and 4. Finding the route to destination is extremely simple. The network topology is saved inside the nodes with a small memory size by using the TMT.

IV. Modeling Per-Packet Energy Consumption Experimental

The sensor consumes energy in receive or send process. The idea of this modelling is by fixing a component that is associated with the device and channel acquisition overhead, proportional incremental to the size of packet. However, such modelling does not consider consumption of energy an unsuccessful try to acquire the channel, or being lost due to collision bit error or loss wireless connectivity. This model works on unicast traffic and unicast mode, and the sensor motes in all traffic by sending nearby motes. The most important thing to consider here is the energy consumption when nodes do not determine the destination of unicast receiver and the back of received packet. The IEEE 802.15.4 standard defines the protocol and interconnection of device via radio communication in personal area network (PAN) called low rate wireless personal area network (LR-WPAN) that uses a carrier sense multiple access with collision avoidance (CSM/CA). However, in this work, the different structures are proposed of CSMA/CA in order to investigate overhead channel acquisition. This structure contains a footer and header data frame, with transfer reliability during packet transmission. In every packet transmission, power consumption is estimated by fixing the component and payload data to make an incremental component of liner equation. For this, a mathematical model will be used to estimate the higher and lower permission routing in our network. An important prerequisite is to carry out this activity in order to develop a methodology for estimating energy consumption in the individual WSN nodes and in the network as a whole. To optimize minimal energy consumption, caring for other parameters such as transmission of parameters should be considered.

In [17], the authors described a linear equation, where the energy is consumed by the sensor node for sending, or receiving the packet. This represents some practical power measurements of CC2420 radio during different operations.

$$\text{Energy} = m \times size + b \qquad (7)$$

Office for Research of Innovation, commercialization and consultancy management (ORICC), (VOT No. U1532).

Where coefficients m and b stand for various operations, m represents the incremental cost and b represents fixed costs [18]. Moreover, the size represents the payload (data packet) by the number of bytes. Therefore, power consumption to send packet is:

$$\text{Energy(send)} = m(send) \times size + b(send) \qquad (8)$$

And for receiving packet is:

$$\text{Energy(receive)} = m(\text{receive}) \times size + b(\text{recieve}) \quad (9)$$

The LR-WPAN defines four frame structures:

A. *Data frame*

Based on the data frame structure as shown in Fig. 5, the length of the physical data packet is 11 bytes + (0 to 20 bytes) + n bytes, where n represents the payload data. Practically, (6 bytes) is needed for addressing purposes. Where, (2-byte) addresses are assigned to the sensor motes, and (2 bytes) source PAN identifier is left empty. Then, (2 byte) destination PAN identifier is assigned. Therefore, the length of the data packet gets 11+6 + n bytes= 17 bytes + n byres.

B. *Acknowledgment frame*

The length of the acknowledgment packet is (11 bytes) based on the acknowledgment frame. However, to send data between two nodes, there are sender Node and receiver Node. The sender Node transmits a 17 +n bytes (data frame) and receive 11 bytes (acknowledgment frame). For the receiver Node, it also needs to receive 17 +n bytes (data frame) and sends 11 bytes (acknowledgment frame). Ignore beacons which are transmitted by the coordinator to provide synchronization services in IEEE 802.15.4. This is followed by calculating power consumption for CC2420 (Single-Chip 2.4 GHz IEEE 802.15.4 Compliant ZigBee) through either providing the synchronization of IEEE 802.15.4 if not used in PAN or transmitting the MAC sub-layer using CSMA/CA as follows: 1-initializing the local back off variables or random back off period.2. Then, clear channel assessment to ensure channel is free. 3. After that, data should be transmitted 4. Finally, the acknowledgment frame should be used.

The first step is not included in the measurement and this first step is an internal step in measuring the upper and the lower bounds of energy consumption of rout. Based on the IEEE802.15.4 framework, there is no listening before receiving the acknowledgement by the data transmitter. Moreover, there is no clear channel assessment before sending the acknowledgment by the receiver of the data frame.

- Clear Channel Assessment CCA and Sending: To measure the power needed for sending operation. It requires a 46-byte packet. Where 28 bytes data payload and 18 bytes header and footer. Therefore, power consumptions to send 1 byte, 11 bytes and 18 bytes are 0.12 mJ, 1.32 mJ and 2.16 mJ, respectively.

- Listening and Receiving: To measure receiving power consumption needed for Listening and Receiving operations, and this requires a 46-byte packet for receiving. There is listening for the duration of 10 ms which is a short periodic receive check before receiving the data. It takes 0.58 mJ. Therefore, power consumptions to receive 1 byte, 11 bytes and 18 bytes are 0.12 mJ, 1.3 mJ and 2.13 mJ, respectively. However, based on previous measurements (8) and (9) are written as follows:

$$\text{Energy (send)} = (0.12) \times n + (3.54)\ mJ \qquad (10)$$

$$\text{Energy (receive)} = (0.12) \times n + (4.03)\ mJ \qquad (11)$$

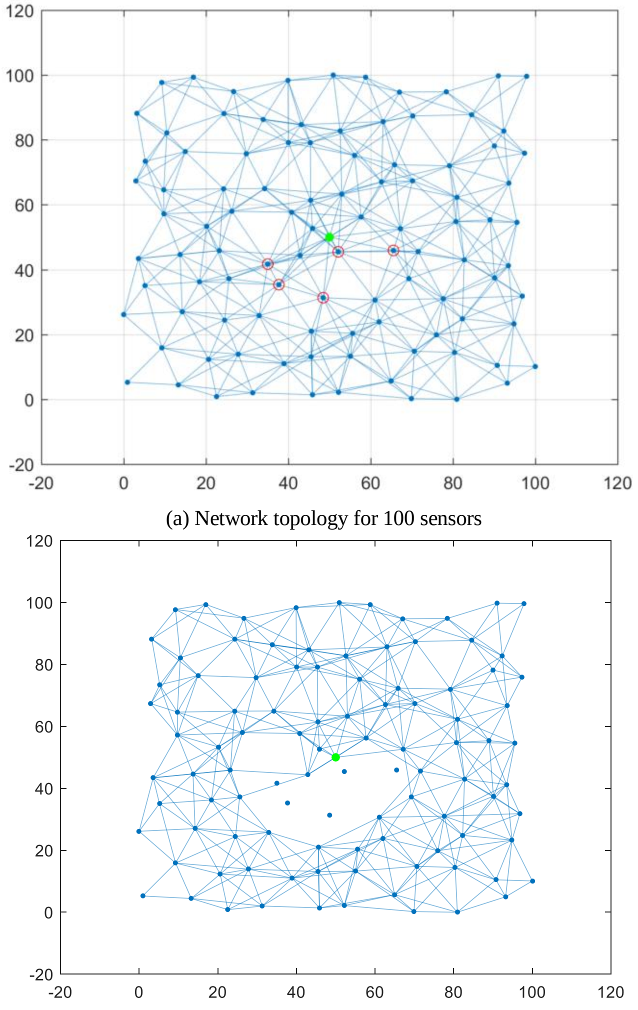
V. CPF Protocol and Packet Energy Consumption

In this section, the routing implementation and how to calculate power consumption for each path in finding the optimal one are explained. In this simulation, using 100 m × 100 m area with 100 Nodes (sensors) scattered randomly. As shown in Fig. 4, the sink node has unlimited power, and it is located in the central area represented by the green color. The simulation parameters are illustrated in Table 1. Basically, sensor nodes use battery such as Nickel Metal Hydride rechargeable. With regular AA, all batteries have a nominal voltage. The charge capacity (C) of battery finishes during time, usually specified as how many amperes a battery can deliver during one hour. For instance, suppose a battery has C = 1200 mAh, this means 1.2 Amperes (1200 mA) for one hour. To measure the total enrage equivalent to the number of Joules, this is performed as follows: Energy (Joules) = Current * 1hour * 3600 sec / 1 hour * V = 1200 mA * 1 hour * 3600 sec / 1 hour * 1.2 V = 5184 Joule. In addition, for each bit, our radio model is assumed to dissipate the energy E_{elec} = 50nJ/bit to run the transmitter or receiver circuit. To transmit the data bits over a distance (d).

TABLE I. SIMULATION PARAMETER

Operation	Parameter	Values
Transmitter/Receiver Electronics	E_{elec}	50nJ/bit
Entail power	E_o	0.5J
Number of bits	k	4000
Number of nodes	n	100
Transmit Amplifier if dmax <= d_0	€fs	10pJ/bit/m2
Transmit Amplifier if dmax >=d_0	€mp	0.0013pJ/bit/m4
area	x, y	100 m × 100 m

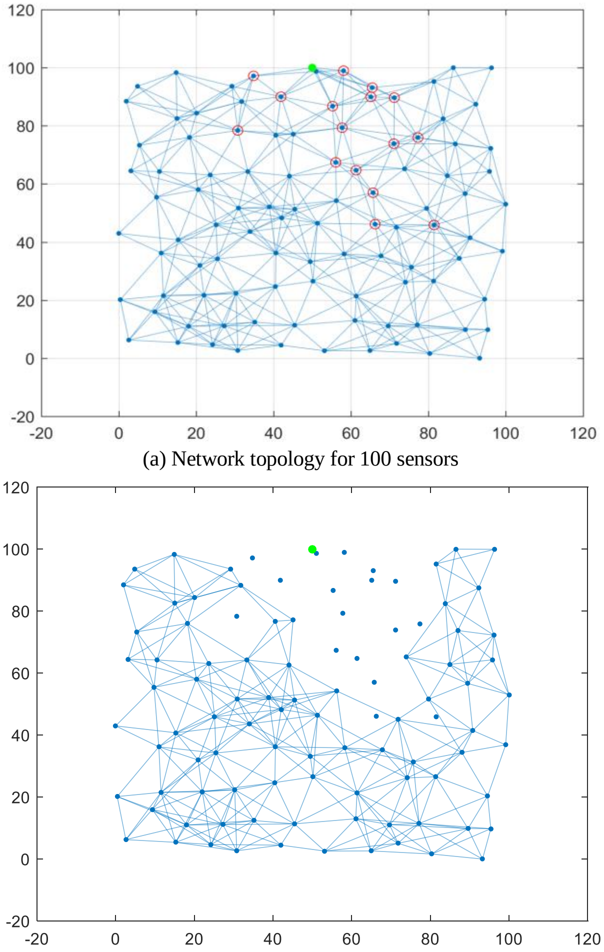
After running the simulation, the first nod dies after 325 iterations of sending the data. Where Node ID is 34. Then, after that, Nodes 35, 36, 44, and 46 die accordingly. Fig. 4 (b) represents the network after 800 iterations of sending the data where the sink node is in the center. Observing the most of nodes die are those nodes which are near to the sink node. This death of such nodes is due to their frequent usage as intermediate nodes when transmitting the data more than the edge nodes. Consequently, they consume more power and die first. In the same experiment, the sink node moves to upper as shown in Fig. 5. The number of dead node increases from 5 nodes in Fig. 4 to 16 nodes in Fig. 5. This occurs due to the same reason mentioned above. It also means that routing protocols do not consume power equally from all nodes in the network. This also happens even in the case of applying energy aware protocols as shown in Fig. 4 and 5.



(a) Network topology for 100 sensors

(b) Network topology with dying nodes

Fig. 4. Network topology for 100 sensors with sink node in the center of field



(a) Network topology for 100 sensors

(b) Network topology with dying nodes

Fig. 5. Network topology for 100 sensors with sink node in the top of field

Because of this problem, the proposed mathematical model can be find the optimal route by consuming power equally from all nodes in the network. The optimal route is proved in the result section by comparing it with previous studies. This resulted in increasing the lifetime of the network. Our new proposed Consume Power Fairly (CPF) protocol method is illustrated below:

1) Adding a Triangular Matrix Table for every node in the network to save the whole network topology information.

2) Using a table to save the power remaining for every node in the network.

3) Finding the possibility of all routes to the destination node (sink) with the help of TMT.

4) Applying our proposed equations to find the best route. As follows:

5) Assuming that the Path consists of several intermediate nodes to the destination. Then, the total power for that route is stated as in (2).

a) The power needed to transmit a particular data is given in (5), which is the power required to transmit a particular size data.

b) (6) provides the total residual power for that particular route, where the result gives several paths.

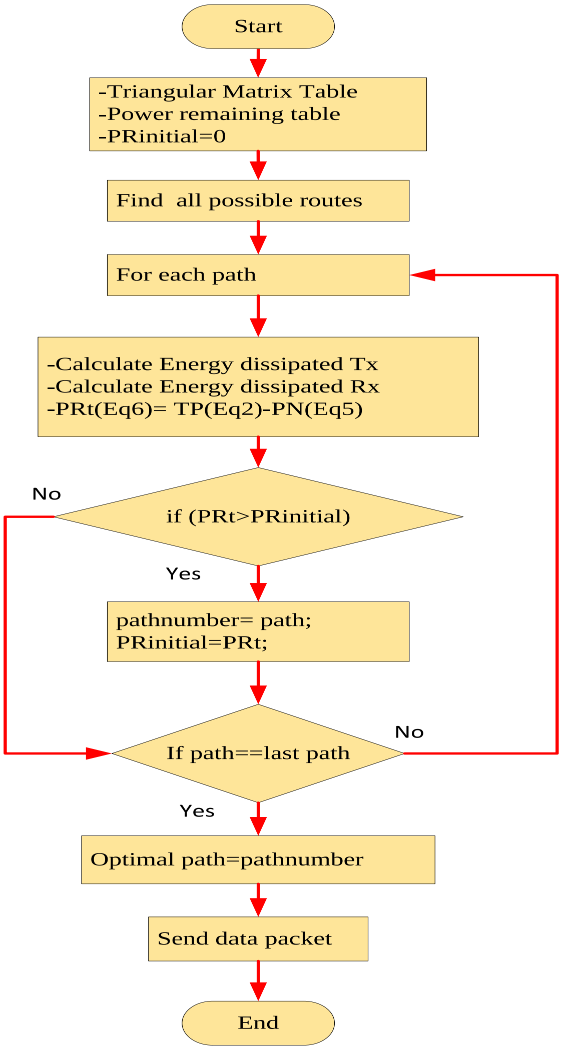
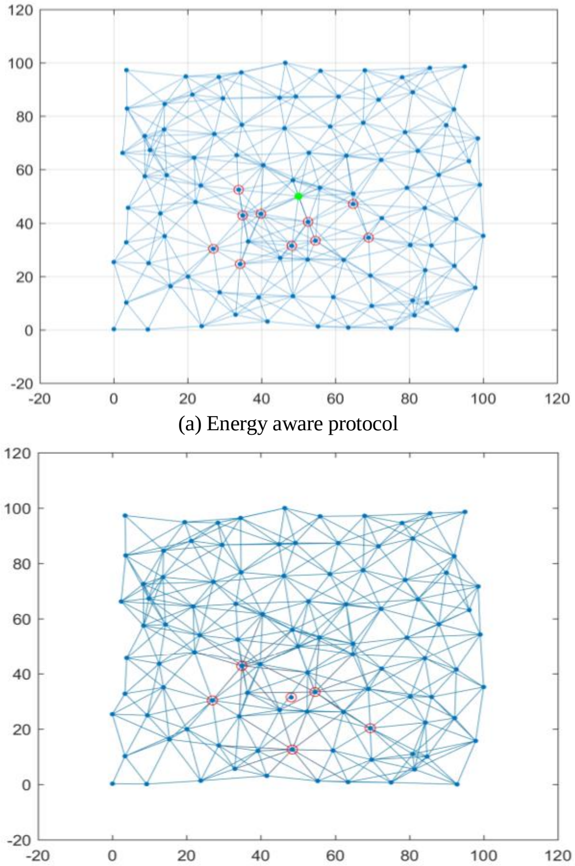


Fig. 6. Method of protocol

However, the optimal route that is selected depends on getting the high number of PR after applying new proposed (6). As a result, prolonging Node Lifetime is increased by applying new Consume Power Fairly (CPF) protocol compared with energy aware protocol as shown in Fig. 7.



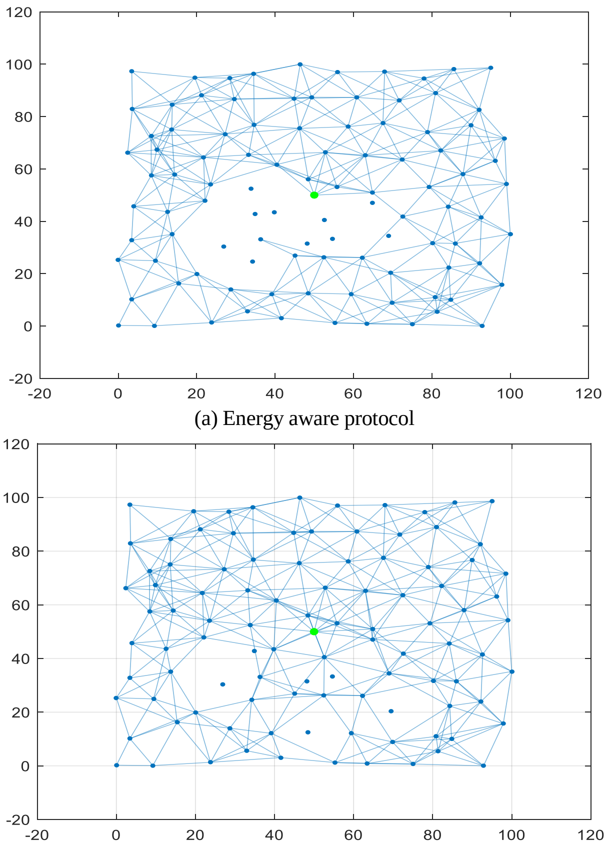
(a) Energy aware protocol

(b) Consume Power Fairly protocol

Fig. 7. Network topology for energy aware protocol compared to CPF protocol

In Fig. 8 (a), the energy aware protocol generates holes of dying nodes that will affect the whole network. Consequently, the network will die fast. In Fig. 8 (b), the number of dying nodes is 6 if compared with 10 dying nodes for energy aware protocol. The iteration of sending data is 1000 number of rounds for both results. Secondly, the dying node distributed in the field will not affect the rest of other nodes where the holes affect the dying node itself only (See Fig. 8-b).

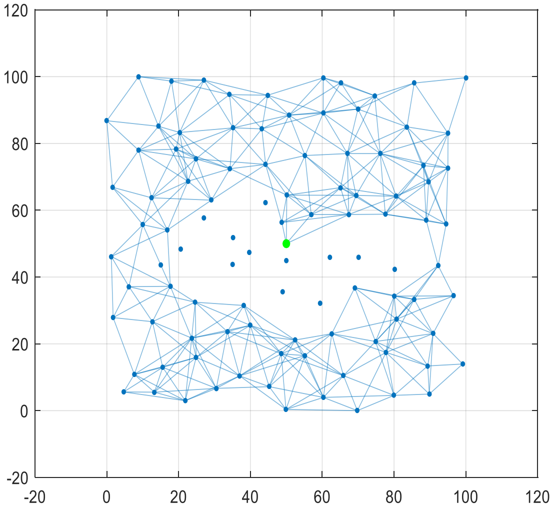
Fig. (8) illustrates how the number of dying nodes affects the network overall. The energy aware protocol generates a big hole around the sink node through 1200 iteration as shown in Fig. 9 (a). This will cause fast death to the whole network. This is because these protocols depend on energy aware protocols that select the optimal path with minimum hops count. In addition, they do not consider the network topology to keep the network live for a long period. Unlike this, our proposed model considers network topology with the help of TMT as shown in (2), (5), and (6) to select the optimal path. Selecting the optimal path may not be the minimum hops count. Therefore, Fig. 9 (b) shows that the network topology is strongly connected, which reflects that the dying nodes do not affect the whole network connections. As a result, there is an increase in the prolonging network Lifetime. Fig. 10 also shows the number of alive nodes. In summary, the proposal CPF protocol reduces the number of dead nodes and keeps the connectivity high, which increases prolonging network Lifetime.



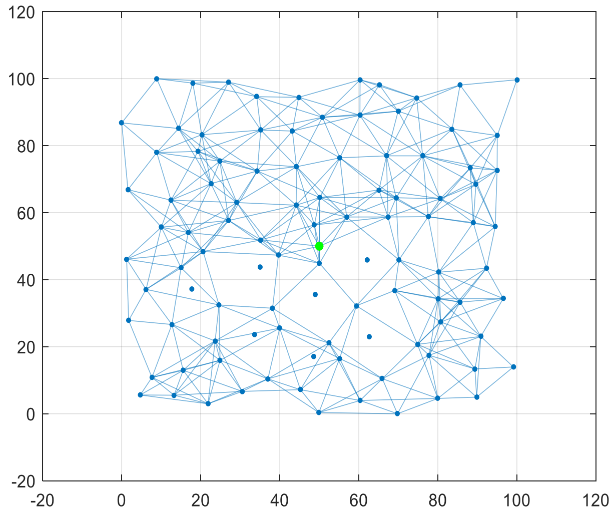
(a) Energy aware protocol

(b) Consume Power Fairly protocol

Fig. 8. Network topology after nodes dying for energy aware protocol compared to CPF protocol for 100 sensors with dying nodes



(a) Energy aware protocol



(b) Consume Power Fairly protocol

Fig. 9. Network topology after 1200 number of iterations for 100 sensors with dying nodes

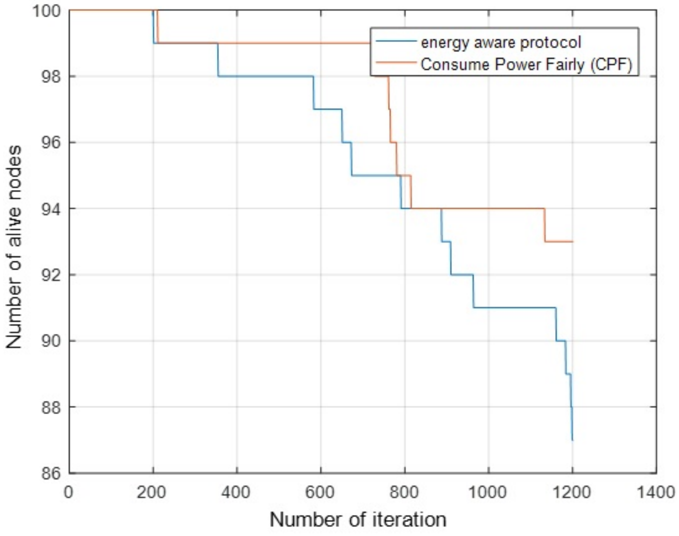


Fig. 10. Number of alive nodes after 1200 iterations

TABLE II. COMPARE BETWEEN TWO PROTOCOLS

Iteration	No. of node die at consume power fairly	No. of node die at energy aware protocol	Consume power fairly network status	Energy aware protocol status
1000	6	10	Connected	Small hole
1200	7	13	Connected	Big hole

VI. Conclusion

The main objective of this paper is to fulfil Prolonging Lifetime for WSN. In achieving this research objective, two contributions to previous research are offered. Firstly, the new mathematical model is proposed to increase the lifetime of WSN. Secondly, the mathematical model was applied to our new protocol called Consume Power Fairly (CPF), where selecting the optimal math depends on consume power fairly among all nodes inside the network to increase the lifetime of the network. An efficient route based on the power consumption and network topology are presented, which can effectively control and distribute the power over the whole WSN to save power for prolonging network lifetime.

Acknowledgment

The authors would like to thank the staff of wireless and radio science center (WARAS) of Universiti Tun Hussein Onn Malaysia (UTHM) for support. Research funding for this work is fully provided by Office for Research of Innovation, commercialization and consultancy management (ORICC), (VOT No. U1532).

References

[1] S. Tanwar, N. Kumar, and J. J. P. C. Rodrigues, “A systematic review on heterogeneous routing protocols for wireless sensor network, ” Journal of Network and Computer Applications, vol. 53, pp. 39-56, 2015.

[2] C. Sergiou, V. Vassiliou, and A. Paphitis, “Congestion control in Wireless Sensor Networks through dynamic alternative path selection, ” Computer Networks, vol. 75, pp. 226-238, 2014.

[3] C. Sergiou, V. Vassiliou, and A. Paphitis, “Congestion control in Wireless Sensor Networks through dynamic alternative path selection, ” Computer Networks, vol. 75, Part A, pp. 226-238, 2014.

[4] A. Razaque and K. Elleithy, “Modular Energy-Efficient and Robust Paradigms for a Disaster-Recovery Process over Wireless Sensor Networks, ” Sensors, vol. 15, no.7, pp.16162-16195, 2015.

[5] L. Liu, N. Zhang, and Y. Liu, “Topology control models and solutions for signal irregularity in mobile underwater wireless sensor networks,” Journal of Network and Computer Applications, vol. 51, pp. 68-90, 2015.

[6] S. Chen, T. Sun, J. Yuan, X. Geng, C. Li, S. Ullah, et al., “Performance analysis of IEEE 802.15.4e Time Slotted Channel Hopping for low-rate wireless networks,” KSII Transactions on Internet and Information Systems, vol. 7, no. 1, pp. 1-21, 2013.

[7] A. A. T. Rahem, M. Ismail, I. A. Najm, and M. Balfaqih, “Topology sense and graph-based TSG: efficient wireless ad hoc routing protocol for WANET,” Telecommunication Systems, pp. 1-16, 2017.

[8] A. Saad, A. J. Hussein, I. A. Najm, and A. T. Rahem, “Vehicular ad hoc networks: Growth and survey for three layers,” Indian Journal of Science and Technology, vol. 7, no. 1, pp. 1-21, 2017.

[9] A. T. Rahem, M. ismail, M. Fadhil, and A. Jamal, “Studying and analyzing algorithm behavior and mechanism for wireless ad hoc routing protocols,” Journal of Engineering and Applied Sciences, vol. 11, no. 19, pp. 11760- 11769, 2016.

[10] A. T. Rahem, M. Ismail, A. Idri, and A. Dheyaa, “A comparative and analysis study of VANET routing protocols,” Journal of Theoretical and Applied Information Technology, vol. 66, no. 3, pp. 691-698, 2014.

[11] A. P. Silva, S. Burleigh, C. M. Hirata, and K. Obraczka, “A survey on congestion control for delay and disruption tolerant networks,” Ad Hoc Networks, vol. 25, Part B, pp. 480-494, 2015.

[12] N. Parrado and Y. Donoso, “Congestion Based Mechanism for Route Discovery in a V2I-V2V System Applying Smart Devices and IoT,” Sensors, vol. 15, no, 4, pp. 7768-7806, 2015.

[13] A. Ghaffari, “Congestion control mechanisms in wireless sensor networks: A survey, ” Journal of Network and Computer Applications, vol. 52, pp. 101-115, 2015.

[14] A. T. Rahem, M. Ismail, N. F. Abdullah, and M. Balfaqih, “Node cooperation to avoid early congestion detection congestion by alternative route based on cross-layer for wireless ad hoc networks,” IJECE, vol. 6, no. 5, pp. 2322-2330, 2016.

[15] R. F. Yezid Donoso, “Multi-Objective Optimization in Computer Networks Using Metaheuristics,” Boca Raton New York: AUERBACH PUBLICATIONS, Dec 12, 2010, Press.

[16] A. T. Rahem, M. Ismail, and A. Saad, “A triangular matrix routing table representation for efficient routing in manet,” Journal of Theoretical & Applied Information Technology, vol. 64, no. 2, pp.401-412, 2014.

[17] L. M. Feeney and M. Nilsson, “Investigating the energy consumption of a wireless network interface in an ad hoc networking environment,” in INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2001, vol. 3, pp. 1548-1557.

[18] M. Amiri, “Evaluation of lifetime bounds of wireless sensor networks,” Computer Research Repository (CoRR), vol. abs/1011.2103, 2010.

Unsupervised Commercials Identification in Videos

Najeed Ahmed Khan
Dept. of Computer Sc. & Software Engineering
NED University of Engineering & Technology
Karachi, Pakistan

Waseemullah
Dept. of Computer Sc. & Software Engineering
NED University of Engineering & Technology
Karachi, Pakistan

Umair Amin
We4do solutions Karachi,
Pakistan

Muhammad Umer
NED University of Engineering & Technology
Karachi, Pakistan

***Abstract*—Commercials (ads) identification and measure their statistics from a video stream is an essential requirement. The duration of a commercial and the timing on which the commercial runs on TV cost differently to the ads owner. Automatic systems that measure these statistics will facilitate the ad owner. This research presents a system that segment semantic videos and identify commercials automatically from broadcast TV transmission. The proposed technique uses color histogram and SURF features resulting in identify individual ads from TV transmission video stream. Experimental results on unseen videos demonstrate better results for ads identification. The target for the proposed approach is television transmission that do not use blank frame between the ads and a non-ad part of the transmission like in Pakistan, different from European countries TV transmission. The proposed segmentation approach is unsupervised.**

Keywords—TV commercial; semantic analysis; segmentation; video classification; commercial detection; commercial classification

I. Introduction

Commercials display on broadcasted TV transmission are a very important part of transmission as majority of revenue for a broadcaster is generated by advertising as well as useful sources of information for the viewers. Knowing what, when and who is advertising can be useful information in knowing market trends and forming business strategy. These commercials can be used as interesting object or segment for semantic analysis of videos.

The term semantics is a very broad and can be used in several different domains e.g. sports, drama, a song, commercial etc. The proposed framework chose commercials that appear in television transmission broadcast for semantic analysis. The target was to develop a framework that can differentiate between commercial and non- commercial segment and compute the statistics of any particular commercial in a TV transmission video stream. The statistics can describe the timing, duration and frequency etc. of a target commercial in a recorded TV transmission video stream. Although several people have attempted commercial detection and identification but most of the work relates to the Europe and USA television transmission which follow transmission standards. Whereas most of the TV channels transmission in Pakistan does not follow those standards e.g. appearance of a blank frame between the ads and a non-ad part of the transmission.

The proposed framework is developed for the transmission that does not use presence of black frames between commercials like in Pakistan TV transmission. Several techniques such as [1] make use presence of black frames between commercials for boundary detection of commercials, which technique is not commonly used for TV channels transmission in Pakistan. Others have used absence of channel logo during the commercial break as a way of separating a video between commercial and non-commercial segments [2] this only technique is again not uses in TV transmission in Pakistan.

This paper has been divided into 7 sections. Section I provides an introduction and overview of the project. Section II discussed related work. Sections III describe Video segmentation implementation; Section IV provides description for commercial identification. Section V explains the method of using Commercials Analysis Application to demonstrate the algorithm. Section VI presents results of experiments that were conducted and Section VII discuss the Future work.

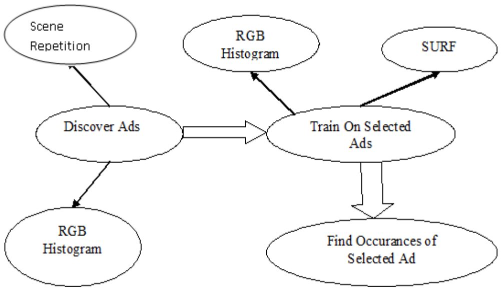


Fig. 1. Framework for proposed technique

II. Literature Review

Most of the work has been done with European or American television transmissions those use presence of black frames between commercials as target for analysis.

In [2] the problem of identifying and categorizing of commercial from TV videos is explored. A multi modal approach is used for boundary detection of individual commercials. Black frames and silence is used for separating ads. Text detection on ads is done for classification of ads. For the purpose of separating commercials from normal transmission absence of channel logo in commercial segment is used. Audio and visual features were used and trained on a Hidden Markov Model. After evaluation, a precision value of 90% and a recall value of 80% were observed.

Covell M. et al in [3] explored the problem of possibility of video segments that are repeated in a video transmission stream. Their major area of concern was the detection of those advertisements which are related to broadcasters own programs such as ads of programs that will be broadcasted next week of on same day at a different time. Their suggested approach has three major steps audio repetition detection, visual descriptors and the endpoint detection. For the purpose of evaluation experiment was run on four days video footage that was captured from different TV channels transmission. A precision rate of 85% and recall rate of 94% was reported for audio matching part. After video matching, precision rate was 92% and recall was 93% reported and in final result precision was 99% and recall was 95% reported.

In [4] the problem of detecting commercials in videos that are encoded in H.264/AVC is explored. This approach is unique in a sense that it works directly on compressed stream instead of having uncompressed video as a unique separate step. The proposed approach makes use of the fact that logo of channel is not present during commercial segment which is true for European and particularly German television. For the purpose of evaluation recordings from 19 television channels was used that were predominantly related German viewers. An average recall value of 97.33% and precision value 99.31% is reported.

In [5] the problem of detection of scene change in video by making use of audio and visual information that is available from video is explored. The proposed method consist of first determining shot boundary detection using and unsupervised segmentation algorithm making use of object tracking. For the purpose of evaluation several videos from TV news were used and after running experiment average recall value of 89% and average precision value of 92% was observed.

In [6] the problem of detecting commercial breaks in MPEG compresses video stream is explored. It makes use of features that are derived from MPEG parameters. For commercial detection presence of black frame, unicolor frame and change in aspect ratio is used. It was also observed that minimum duration of commercial break is one minute. For the purpose of evaluation television transmission of eight hour length was recorded from Dutch TV stations and presence of black frames in commercial break was found to be the strongest ad detecting parameter.

In [7] a learning based approach for the detection of TV commercials is proposed. Their approach is to do Support Vector Machine based classification that is based on several visual and audio features. Used visual features include average of Edge Change Ration, variance of edge change ratio, average of frame difference and variance of frame difference. Melfrequency Cepstral Coefficient and short time energy are used as audio features. Some post processing steps such as removing of scenes that have very small length, checking of long commercials and refining of commercial boundaries. For purpose of evaluation 10.75 hour of recording TV transmission was used that was collected from different TV channels such as NBC, ESPN2 and CNN. Without post processing a recall value of 88.21% was observed that increased to 91.77% and a precision value of 89.39% was observed without applying post processing and when post processing was applied, it increased to 91.65%

[8] has proposed a method for automatic unsupervised segmentation of TV content that makes use of a signal based approach. This approach can be applied to audio, visual or a combination of audio and visual signal by making use of general likelihood ratio and Bayesian Information Criterion .This system was evaluated on recordings from French television and also on TRECVid dataset. For ARGO as recall value of 93% was observed whereas for TRECVid recall value of 89% was observed. For ARGO precision value of 93% was observed whereas for TRECVid precision value of 91% was observed.

[9] has proposed a novel method for commercial detection that is centered around cookery programs. Commercial boundaries are detected based on presence of audiovisual features. Initially different audio features are used for detecting the start and end of commercial break. Then logo of program name is matched with start and end of commercial break. Zero crossing rate and short time energy are used as audio features. Edge detection and corner detection is used for visual analysis

In [10] the problem of automatically annotating broadcast videos for later search and indexing is discussed as manual annotating is a very costly time consuming and subjective. The approach used is to apply multi-modal machine learning techniques to audio video and text components of video for analysis and retrieval. The system is unique in that it uses and combines audio, video and text information present in video for annotation. Machine learning is applied to create a library of semantic models from training dataset. Human interaction is required in training phase but just for a small dataset. The system allowed users to query videos in several ways based either on feature by selecting a key frame, text based, semantic based or based on model. For the purpose of evaluation TRECVID benchmark was used. Accuracy of 90% was achieved.

In [11] the problem of real-time indexing of videos based on their content is investigated. The suggested approach is to apply statistical methods using Hidden Markov Model (HMM) for content based video indexing. Most of the features used as input to HMM are based of difference image sequence that specify the motion of main object in scene. Other used features are average motion deviation that helps to distinguish shots where large parts are in motion, grey level histogram that is useful in detection of cuts, center of motion and overall intensity of motion. This approach merge scene detection and scene classification in a single step because having scene detection and scene classification as two separate steps cause a fault in scene detection to result in wrong classification. For evaluation recording of 12 news shows recorded from different German TV stations were used. Six news shows were used for training and six were used for testing. They were classified into nine classes namely Studio Speaker, Report, Begin, End, Weather Forecast, Out, Interview , Cut ,Frame Translate and Window Change. Out of nine classes seven had recognition rate greater than 80%. It was found that recognition rate for short news is significantly better than that for long news.

In [12] the problem of classifying feature films into categories based on their preview is explored. They have classified films into four categories namely Comedy, Action, Drama and Horror based only on computable visual cues. The suggested approach is to describe input as a set of features that are likely to minimize variance of points within a class and maximize variance between points of different class. The features used included average shot length that was computed by using average color histogram in HSV (Hue, Saturation, Value of brightness) space. For the purpose of evaluation 101 movie previews were taken from apple website and classified into four categories. Out of 101 movies 17 were wrongly classified.

In [13] the problem of retrieving commercial stream based on their salient semantics is discussed from a semiotic perspective. Four semiotic categories of commercials were identified namely practical, playful, utopic and critical. For evaluation purpose 150 commercials from several Italian channels were used and tests were conducted to verify that classifications done by system are in conformity with that done by human experts. The best results were seen for playful commercials and worst performance was for practical features and reason was inability of system to properly detect that the promoted product is in foreground or not.

III. Video Segmentation

The first step of the proposed framework is the video segmentation based on semantics [14]. These segments will broadly consist of different programs in addition to the commercial ads. The target is to segment commercials and then identify any particular commercial in the video stream. In order to automatically segment semantic videos, RGB mean is computed for all frames in video, plot histogram corresponding to the frame numbers and calculate variance. From the plots, it is observed that there was a pattern being followed by a commercial segment and normal transmission. The commercial segment had higher distortion in graph whereas the segment that corresponds to normal transmission had fewer peaks per second. Based on histogram variance information the video is segmented in to commercial and no-commercial segments. The details of video segmentation based on semantic can be reviewed in [15].

From the segmented video the first step is to detect start and stop boundary of a commercial.

Ad Boundary Detection

After segmenting video into commercial segment and non-commercial segment the aim is to find automatically the boundaries of individual commercials. Mostly the existing systems [15] make use of property of black frame existence at the end of each commercial. But the black frame existence was not present in all countries TV transmission channels, like in Pakistan. Therefore, a technique for automatically detection of commercials in such kind of TV transmission is needed.

To detect boundary of a commercial in video stream, color histogram of repeating patterns are computed, as the commercials are usually repeated several times during transmission. In order to determine a frame belong to a same scene or not, we calculate histogram difference of change point with last 5 scenes, if a match was found then the frame belongs to the existing scene.

After a scene is identified it is compared with other scenes those were already been generated from the video. If a new scene is found that was not present before it is assigned a new scene ID. However if that scene was already been observed in an earlier location in the video it is added to the list of scenes belonging to existing scene ID.

In the next step repeating segments of scenes are computed. This allows detection of unique commercials because normally a commercial segment is composed of several scenes and it appears several times in a TV transmission as compared with other programs in transmission. The details of the algorithm can be reviewed in [14]. Each discovered pattern generally represents a commercial segment that was repeated several time in the video

IV. Commercial Identification

Commercials identification is the step where the location(s) and duration(s) of a particular commercial if exist in the video stream are investigated. It uses a trained feature file as reference for target commercial. This section describes training and commercial identification phases.

- Training

In the Training phase first select frames from a specific commercial for which the system is required to train and RGB histogram is calculated. Using histogram variance information the commercial segment is split into scenes. If the histogram variance of two consecutive frames is greater than threshold the point is considered as a change point for scene. Then in second step user asks to select a *key_object* in the commercial segment which is significant with respect to the product of commercial. The *key_object* is used for computation of SURF (Speed Up Robust Transform) feature. Example of a selected key object is shown in figure 2. In this figure a bottle of Lifebuoy shampoo is selected as key_object from a video frame.

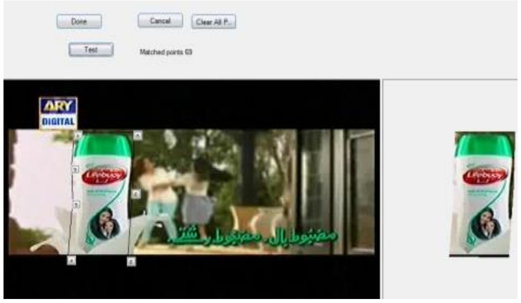


Fig. 2. Selection of object for SURF feature computation

One or more key_objects can also be selected for computation of SURF feature; however, selecting one object is enough to eliminate false positive ad detections. After this the computed SURF feature file uses as a training file for that specific commercial, which is used for the commercial identification stage in the transmission video stream.

- Identification

In the identification phase the training file is uses that find the presence of the investigated commercial segments in the provided video stream. It finds all possible matches (if exist) of the key_object in the video and mark all related frames in the video stream. Sometime a long commercial segment is broadcasted first time and after some time a shorter version of that commercial is broadcast. In this case it is recommended that the system should train on longer version of commercial. This will also be able to detect or identified shorter version of the commercials. To reduce the search space, histogram of each frame in the scene is compared with only the histogram of first frame in the trained scene. The details of commercials identification step is given below.

First histogram of each frames is matched with histogram of previous frame and if the difference is greater than the threshold a new scene is generated. For each new generated scene, histogram of all frames is compared with histogram of first frame the trained scene. If it matches the criteria for 2/3rd match with any one of those scenes then the scene is kept as candidate for the investigated commercial otherwise it is ignored and move to a next scene. When a frame is found where histogram difference is less than the threshold for the object scene then the content of that frame is read and first generate Integral Image from RGB image. Next interest points are calculated for current frame, and then matches are determined between described interest points of object and described interest points of frame. If number of matched interest points is greater than 10 (set a threshold) current commercial segment keep as a valid commercial otherwise it is considered as false detection and deleted.

The advantage of this approach is that we do not need all frames data of video transmission for investigating a commercial because only the histogram information is need that can be computed once only. Frame content is only needed for training stage to compute SURF descriptors of key_object of the commercial. Therefore, new training files can easily be added to the system for new commercial to be investigated. For example if we have a training file of e.g. ‘DEW’ commercial that will detect commercial segments of DEW appear in transmission stream and we get asked to find occurrences of other commercial e.g. “Fair and Lovely” in a the transmission stream, all we need to do is make a training file for Fair and Lovely commercial by computing SURF descriptors of Fair and Lovely key_object and run it on system. Because histograms have already been calculated when we checked for DEW so the system will just load them from file and can give result quickly about presence of Fair and Lovely commercial.

V. COMMERCIALS ANALYSIS APPLICATION

An application was created to demonstrate how the proposed algorithm can be used for detection, identification and computing statistics of a target commercial in a video stream. It consists of four (4) main modules:

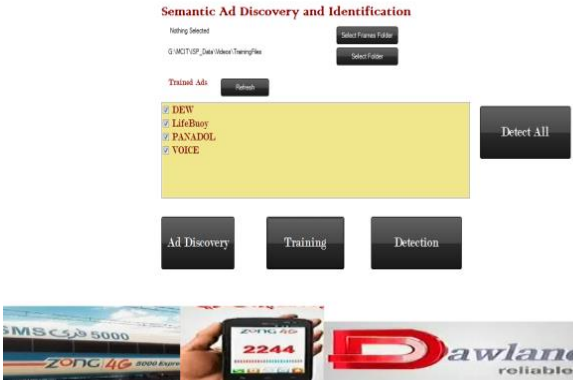
Commercial Discovery is detection of all commercial segments at all positions where they are present inside a video stream based on Scene Identification and detecting repeating patterns of scenes.

Training is used to train the system for a new commercial that was previously unknown to the system.

Detection is used for detection or identification and verification of single ad based on training file that is provided.

Detect All allows user to check for presence of all commercials that are known to the system in a specified video stream and provide user with a summary result.

The screen shot of the designed application shown in figure 3. The details of each step are given below.



Semantic Ad Discovery and Identification

Nothing Selected

Select Frames Folder

G:\MCIT\SP_Data\Videos\TrainingFiles

Select Folder

Trained Ads

Refresh

☑ DEW
☑ LifeBuoy
☑ PANADOL
☑ VOICE

Detect All

Ad Discovery

Training

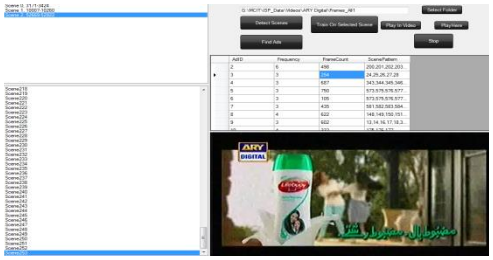
Detection

Fig. 3. Screenshot of main screen, showing options for selecting frames folder and training files for Ad detection

- Commercial Discovery

This part is used for finding unique ads inside the given video based on repeating patterns. At video stream and computed RGB histogram is selected. Clicking on Detect Scenes button extracts scenes from frames that present in the selected folder. Find Ads will apply repetition detection algorithm to look for repeating patterns of scenes and mark them as unique ads.

Result can be viewed in a grid and any row can be selected to generate a training data for that row. Figure 4 shows screen Commercial discovery grid, from which any desired ad (scene) may be selected.



Scene 0: 3171-3424
Scene 1: 10007-10260
Scene 2: [illegible]

G:\MCIT\SP_Data\Videos\ARY Digital\Frames_All

Select Folder

Detect Scenes

Train On Selected Scene

Play In Video

PlayHere

Find Ads

Stop

AdID	Frequency	FrameCount	ScenePattern
2	6	498	200,201,202,203...
3	3	254	24,25,26,27,28
4	3	687	343,344,345,346...
5	3	750	573,575,576,577...
6	3	105	573,575,576,577...
7	3	435	581,582,583,584...
8	4	622	148,149,150,151...
9	3	602	13,14,16,17,18,3...

Scene218
Scene219
Scene220
Scene221
Scene222
Scene223
Scene224
Scene225
Scene226
Scene227
Scene228
Scene229
Scene230
Scene231
Scene232
Scene233
Scene234
Scene235
Scene236
Scene237
Scene238
Scene239
Scene240
Scene241
Scene242
Scene243
Scene244
Scene245
Scene246
Scene247
Scene248
Scene249
Scene250
Scene251

Fig. 4. Commercial Discovery Screen

- Training

In this screen user can view the selected scenes for a selected commercial and view the parameters that are detected for selected scene or frame. From the selected scene user can click on "Select SURF Object" to select object that will be used for computation of SURF features.

Following three parameters are to be computed.

- MaxDiff is maximum histogram difference between start frame and any other frame of scene.
- LastDiff is value of histogram difference between current frame and previous frame.
- StartDiff is value of histogram difference between current frame and first frame of scene.

- Object Selector

Object selector screen allows user to select a key-object that is used for calculating reference SURF descriptor points. In figure 5, on the left side complete image is shown and on the right side selected object is showed. Clicking on "Test" button calculates surf descriptors for selected region and also calculates surf descriptor for complete image. Then these two are compared to get matching surf points. Number of matching surf points are shown to the user which can be used to decide if selected object is a good candidate for SURF matching or a different object may be chosen.

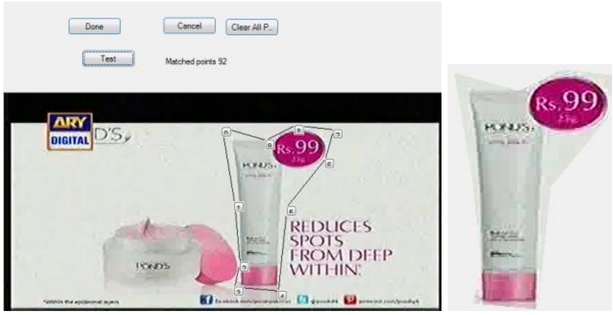


Fig. 5. Object Selector Screen**,** left side complete image right side selected cropped object of interest

- Commercial Detection

Individual commercial detection steps can be performed by clicking on buttons "Detection Steps" or all steps can be performed by clicking on "Perform Complete Detection" that will perform detection and will show results single ad as shown in figure 6 and all ads shown in figure 7 respectively by indicating a timeline with commercial region highlighted.

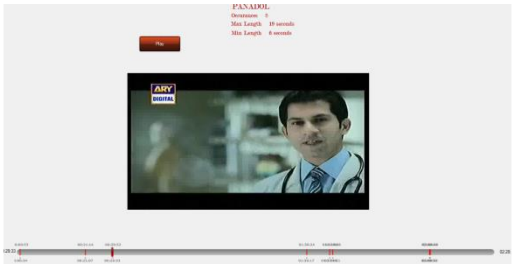


Fig. 6. Single Commercial Detection

In figure 6, start time and end time is shown by the timeline with *red color bar* highlighted regions. This figure shows detection of commercial for Panadol, which was appeared 8 times in total in the video stream. Minimum length of that commercial in a single instance was 6 seconds and maximum length of that commercial in another instance was 19 seconds.

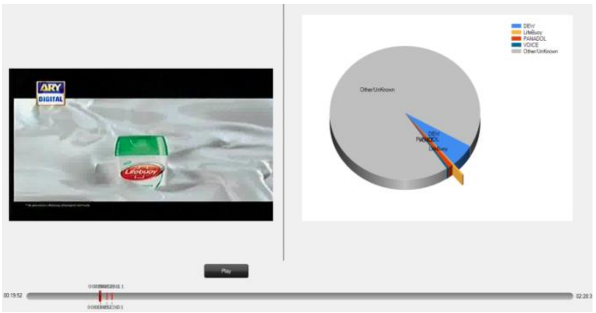


Fig. 7. Multiple Commercials Identification (Lifebuoy Shampoo selected)

Figure 7 shows multiple commercials detected include DEW, Lifebuoy, PANADOL, VOICE and Other Unknown in the input video shown in pi graph and timeline. Each region can be clicked to view the location and size of that commercial in timeline of original provided video.

VI. Results

This section will describe results and statistics by running our algorithm on the test dataset. Test dataset composed of 03 segments of transmission that was recorded from two different TV channels - ARY Digital and Hum TV. Two segments of ARY digital were 2 hour 30 minutes in length and HUM TV segment was 1 hour 40 minutes in length. All 3 segments were recorded at 25 frames per second.

For evaluation, the commercials in all 3 recorded segments were first marked manually and then the detection algorithm was executed. Outcome automatic detection results were compared with manually marked commercials. Target in this test was to find those ads that appear at least 3 times in the test video segment. Detection results for 3 test video segments are given in the Tables.

Table1 shows results for video segment recorded from HUM TV channel.

TABLE I. Detection Results for Commercial Segment in Video Recorded from HUM TV Channel

Commercial	Present Duration (Sec)	Detected Duration(Sec)
Care	200	170
Dawn	45	0
Express	15	15
Nestle	45	0
Stillmens	15	15
Express	15	15
Tapal	30	30
Nido	120	120
Total	485	365

For segment from HUM television there were no false positive results.

Table 2, shows results for video segment recorded from ARY digital TV channel.

TABLE II. DETECTION RESULTS FOR COMMERCIAL SEGMENT IN VIDEO RECORDED FROM ARY DIGITAL CHANNEL

Commercial	Present Duration (Sec)	Detected Duration(Sec)
Dairy Omung	200	183
Dawlance	45	45
Dettol	30	30
Dew	480	480
Dove	90	90
Fair&Lovely	40	0
Horlicks	126	126
LifeBuoy	30	30
Olpers	60	60
Panadol	70	65
Sensodyne	90	70
Veet	60	60
Zong	80	80
Total	1401	1325

Third data segment is another different 2 hour 30 minute recording of ARY digital. Results for that dataset are presented in Table 3.

TABLE III. DETECTION RESULTS FOR COMMERCIAL SEGMENT IN 2^{ND} VIDEO RECORDED FROM ARY DIGITAL CHANNEL

Commercial	Present Duration (Sec)	Detected Duration(Sec)
Dove	95	85
Telenor	55	55
Ponds	120	120
Knor	165	165
Head&Shoulders	55	45
LifeBuoy	110	110
Safeguard	44	33
Fair&Lovely	105	105
LifeBuoy	30	30
Dove	100	100
Gluco	45	0
Samsung	15	0
Total	949	848

In total 20 second of invalid commercials content was marked as commercials.

The performance of the algorithm is measured using standard Precision, Recall and F1 value [15]. Statistics of Precision, Recall and F1 value is given in Table 4.

Precision recall and F1 value of each segment is given in table 4.

TABLE IV. IN TABLE, COLUMN T.P IS TOTAL NUMBER OF TRUE POSITIVE DURATION IN SECONDS FOR ADS, F.P ARE TOTAL NUMBER OF FALSE POSITIVE DURATION IN SECONDS, F.N IS TOTAL NUMBER OF FALSE NEGATIVE DURATION IN SECONDS

Segment	T.P	F.P	F.N	Precision	Recall	F1-Value
HUMTV (1)	365	0	129	100%	73%	84%
ARY Digital(1)	1325	50	116	96%	92%	93%
ARY Digital(2)	848	20	101	97%	89%	92%

From table 4, the average Precision, Recall and F-1 values for all three segments are 97.6%, 84.6% and F1 value is 89% respectively.

For evaluation of commercials identification the system was trained on 6 commercials for different products and evaluated. The identification results are given in table 5.

TABLE V. SHOW RESULTS OF AD IDENTIFICATION, AD COLUMN IS NAME OF AD, ACTUAL IS TOTAL NUMBER OF TIMES THAT AD WAS SEEN, DETECTED IS NUMBER OF TIMES THAT AD WAS DETECTED BY SYSTEM

AD	Actual	Detected
DEW	32	32
Lifebuoy	3	3
Panadol	7	7
Safeguard	3	0
Olpers	8	8
Ponds	0 (ad was not present in the video stream)	0

VII. CONCLUSION AND FUTURE WORK

We have developed a framework for commercials detection and identification using color histogram and SURF featured descriptors. The framework performs well for TV transmission having no use of black frame between different programs. The proposed technique is unsupervised and able to differentiate between commercials and noncommercial program parts of transmission in addition to find a way of identifying position, frequency and duration of commercials in a TV transmission stream.

The proposed method can be extended and converted into a complete media monitoring and ad verification package suitable for environment where other solutions are not suitable with a feedback approach used to improve performance of automatic commercial detection. From experiments, it is observed that by increasing the duration of video stream that is analyze, will achieve better results.

REFERENCES

[1] R. Lienhart, C. Kuhmunch, and W. Effelsberg, "On the detection and recognition of television commercials," in *Multimedia Computing and Systems' 97. Proceedings., IEEE International Conference on*, 1997, pp. 509-516.

[2] L.-Y. Duan, J. Wang, Y. Zheng, J. S. Jin, H. Lu, and C. Xu, "Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis," in *Proceedings of the 14th ACM international conference on Multimedia*, 2006, pp. 201-210.

[3] M. Covell, S. Baluja, and M. Fink, "Advertisement detection and replacement using acoustic and visual repetition," in *2006 IEEE Workshop on Multimedia Signal Processing*, 2006, pp. 461-466.

[4] K. Schöffmann, M. Lux, and L. Böszörmenyi, "A novel approach for fast and accurate commercial detection in H. 264/AVC bit streams based on logo identification," in *International Conference on Multimedia Modeling*, 2009, pp. 119-127.

[5] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang, "Scene change detection by audio and video clues," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, 2002, pp. 365-368.

[6] N. Dimitrova, S. Jeannin, J. Nesvadba, T. McGee, L. Agnihotri, and G. Mekenkamp, "Real time commercial detection using MPEG features," in *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowlwdge-based Systems (IPMU2002)*, 2002, pp. 481-486.

[7] X.-S. Hua, L. Lu, and H.-J. Zhang, "Robust learning-based TV commercial detection," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, p. 4 pp.

[8] El-Khoury, C. Sénac, and P. Joly, "Unsupervised segmentation methods of TV contents," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, 2010.

[9] N. Venkatesh, B. Rajeev, and M. G. Chandra, "Novel TV commercial detection in cookery program videos," in *Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II, WCECS 2009*, 2009, pp. 20-22.

[10] J. R. Smith, M. Campbell, M. Naphade, A. Natsev, and J. Tesic, "Learning and classification of semantic concepts in broadcast video," in *Proceedings of the International Conference of Intelligence Analysis*, 2005.

[11] S. Eickeler, A. Kosmala, and G. Rigoll, "A new approach to content-based video indexing using hidden markov models," in *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 1997, pp. 149-154.

[12] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 52-64, 2005.

[13] C. Colombo, A. Del Bimbo, and P. Pala, "Retrieval of commercials by semantic content: the semiotic perspective," *Multimedia Tools and Applications*, vol. 13, pp. 93-118, 2001.

[14] K.A. Najeed, A. Umair, H. Shams, Waseemullah, M. Umer, Unsupervised Ads detection in videos, (accepted for) International Journal of Computer Science and Information Security (IJCISS) *Vol: 15, No.1, January 2017.*

[15] K.A. Najeed, A. Saman, Z. Shehnila, A.Yusra, S. Sehrish, S.H. Saba. Unsupervised Cancer Detection using Computer Vision Techniques, International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 10, pp. 724-732, October 2016.

Computational Model for the Generalised Dispersion of Synovial Fluid

M. Alshehri
College of Computer and Information Sciences
Majmaah University, Saudi Arabia

S. K. Sharma
College of Computer and Information Sciences
Majmaah University, Saudi Arabia

***Abstract*—The metabolic function of synovial fluid is important to understand normal and abnormal synovial joint motion, especially if one seeks some leading causes of the degenerative joint disease. The concentration of hyaluronic acid molecules and other high molecular weight substances in the synovial fluid may be responsible to disperse the nutrients into the cartilage. The theoretical study of the convective diffusion mechanism occurring in the knee joint is presented. A flow model has been analyzed for better understanding of the convective diffusion of the viscous flow in between the articular surfaces. The governing system of partial differential equations has been solved for the Newtonian fluid with suitable matching and conditions. The analytical solution for the unsteady dispersion problem has been obtained for the better understand the phenomena of nutritional transport to synovial joint. The contributions of (convection + diffusion) on the dispersion of nutrients are investigated in detail. The dispersion coefficient has been computed for different values of the viscosity parameter. The results show that the average concentration has a negative correlation with the axial distance and the time.**

Keywords—Synovial Fluid; Articular Cartilage; Unsteady diffusion coefficient; Computational model

I. INTRODUCTION

The synovial joints play a very important role in humans and animal locomotion. These joints sustain the very high loads and low friction under normal physiological condition [1]. Articular cartilage can be considered as a porous gel of proteoglycan aggregates embedded in a water swollen network of collagen fibrils. When cartilage is compressed, its interstitial fluid is forced to flow relative to the solid organic matrix and to be exuded from this matrix [2, 3]. While immature articular cartilage contains vessels that transport the nutrients. The extra-cellular fluid through diffusion and convection [4] transports nutrients from the synovial fluid [5]. The process of dispersion plays a very important role in chemical as well as in biological systems [6-11].

The generalized dispersion model is very useful and valid for all time, for examples of biomedical engineering, namely coronary artery diseases [CAD] and synovial joints [12]. In CAD The suspended particles may execute microcirculation in dispersing through the endothelium. Similarly, with synovial joints, hyaluronic acid [HA], glycoprotein, and other macromolecular components disperse from synovial fluid to cartilage. The diffusion coefficient of hyaluronan in healthy synovial fluid was on average 30% slower than expected by sample viscosity [13]. HA and other components may also execute microcirculation while dispersing from synovial fluid to cartilage. The endothelium in arteries and cartilage in synovial joints are layers of porous cells that may or may not deform.

Synovial fluid plays the key role in the lubrication of the joints, and also for the provision of nutrients and removal of metabolites from the avascular articular cartilage. Boosted lubrication, the process of imbibition and exudation increase the concentration of the hyaluronic acid molecules in synovial fluid [14]. The increased concentration of hyaluronic acid will give rise to the increase in the viscosity of the synovial fluid according to the Wegamirs findings. The macroscopic behavior of the particulate suspension can thus approximate to the homogeneous fluid of greater viscosity than the suspending medium [15]. The artificial joints functioning depend on the dispersion of hyaluronic acid and nutritional transport from synovial fluid to the joints [23-25].

Various attempts have been made by the researchers to investigate the characteristics of dispersion in fluid dynamical situations. Taylor studied under effect the real-time action of molecular diffusion and variation of the velocity of solvent on the dispersion of solute [18]. Gill et al [26] has been obtained the exact solution for the dispersion coefficients. Aris et al [19] study the dispersion process under restricted parameter describing the concentration of solute in terms of its moments in the direction of flow.

In this paper, an effort has been constituted to examine the generalized dispersion of hyaluronic acid particles and other proteins of synovial fluid for the endurance of the cartilage. The joint replacement leads into the rheologically modified lubricant, called periprosthetic fluid, and behaves almost as water low viscosity, Newtonian fluid [20]. It advises us to represent synovial fluid by the viscous fluid in between two approaching rigid plane surfaces. Thus, to know whether an artificial joint works efficiently or not it is essential to look into the dispersion phenomena using Newtonian lubricant. At the porous boundary, Beavers and Joseph boundary conditions with a slightly modified form have been used. For the viscosity of the intermission, the dispersion coefficient is found from the diffusion equation using the generalized hypothesis. The dominant dispersion coefficients for mean concentration have been analyzed in details. It has been set up that the viscosity coefficient decreases as the diffusion coefficient increases. The effects have also been obtained for mean concentration distribution of several values of viscosity. It is observed that viscosity decreases with the mean concentration distribution and increases with the diffusion coefficient.

II. Mathematical Formulation and Solution

The knee joint plays a very important role in human locomotion. It plays an essential for the movement of body as well as carries the many times of the body weight in the horizontally and vertically direction during running and walking almost frictionless. The knee joint is one of the largest and most complex joints in the body. The knee joins the thighbone (femur) to the shinbone (tibia). Tendons connect the knee bones to the leg muscles that move the knee joint. Ligaments join the knee bones and provide stability to the knee in preventive and self-corrective ways. The anterior cruciate ligament prevents the femur from sliding backward on the tibia (or the tibia sliding forward on the femur).The posterior cruciate ligament prevents the femur from sliding forward on the tibia (or the tibia from sliding backward on the femur). The medial and lateral collateral ligaments prevent the femur from sliding side to side.

The configuration of bearing model for analysis consists of two rectangular plates of infinite length (not shown in the figure) in x direction. The surfaces are kept apart by a fluid film of thickness 2h. Introducing the usual assumption of lubrication theory in the Navier–Stokes equation of motion and neglecting the variation of pressure normal to very thin film of lubrication, the following differential equations are obtained for pressure (p) in the fluid film region.

$$-\frac{\partial p}{\partial x}+\mu\frac{\partial^2 u}{\partial y^2}=0 \qquad (1)$$

$$0=-\frac{\partial p}{\partial y}$$

Boundary Conditions:

$$u=-\frac{\sqrt{\emptyset}}{\alpha}\frac{\partial u}{\partial y}\quad at\ y=h$$

$$u=0\qquad at\ y=0 \qquad (2)$$

$$p=0\ at\ x=\pm\frac{x_s}{2}$$

where, $\emptyset$, $\alpha\ and\ \mu$ are porosity, slip parameter and viscosity coefficient respectively

$$\text{Using } \eta=y/h,\ \frac{dp}{dx}=\frac{\mu}{h}\frac{d^2u}{d\eta^2} \qquad (3)$$

$$u=\frac{dp}{dx}\left[\frac{h\eta^2}{2\mu}-\eta\left(\frac{h}{2\mu}+\frac{\sqrt{\emptyset}}{\alpha}\frac{h}{\mu}\right)/\left(1+\frac{\sqrt{\emptyset}}{\alpha h}\right)\right] \qquad (4)$$

$$u^*=\frac{u-\bar{u}}{\bar{u}}=3\eta^2-6\eta\beta-1 \qquad (5)$$

where, $\beta=\left(\frac{1}{2}+\sigma\right)\left(\frac{\alpha}{\alpha+\sigma}\right),\sigma=\frac{\sqrt{\emptyset}}{\alpha}$, u^* is the average velocity in fluid film region.

A simple mass balance between changes in concentration $c(t,x,y)$ of solutes like HA molecules in synovial fluid by convection and diffusion leads to in terms of the dispersion coefficient of making molecules in synovial fluid, got here as approximately spatially uniform.

$$\frac{\partial c}{\partial t}+(u-\bar{u})\frac{\partial c}{\partial x}=D\left(\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}\right) \qquad (6)$$

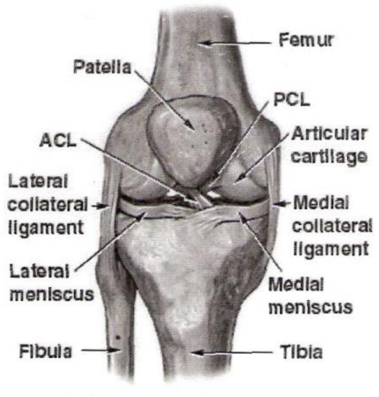


Fig. 1. Right Knee – Anterior view with Patella Tendon Removed

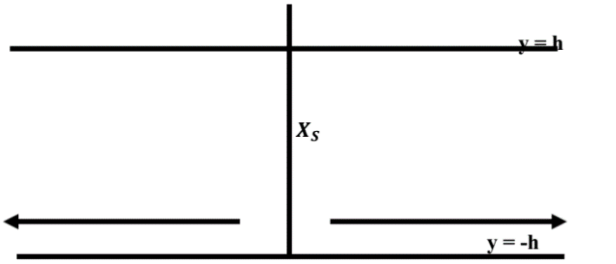


Fig. 2. Parallel Plate Geometry for Dispersion in a synovial fluid

Where $c\ (t,x,y)$ is the concentration of the initial input of length x_s

It has been assumed that the solution concentration is c_0 at the time when the process of imbibition and exudation starts and the superficial and deep layer of articular cartilage is solute free.

$$c(0,x,y)=\ c_0\ \text{ for } |x|\le\frac{1}{2}x_s$$

$$c(0,x,y)=\ 0\ \text{ for } |x|\ge\frac{1}{2}x_s \qquad (7)$$

$$c(t,\infty,y)=0,\ \frac{\partial c}{\partial y}(\tau,x,y)=0\ at\quad y=\pm h$$

and putting in the accompanying non-dimensional system

$$\Theta=\frac{c}{c_0},\xi=\frac{Dx}{h^2\bar{u}},X_s=\frac{Dx_s}{h^2\bar{u}},\eta=\frac{y}{h},\tau=\frac{Dt}{h^2} \qquad (8)$$

The Eqn. (6) and (8) may be written in non-dimensional form as:

$$\frac{\partial\Theta}{\partial\tau}+u^*\frac{\partial\Theta}{\partial\xi}=\frac{1}{{P_e}^2}\frac{\partial^2\Theta}{\partial\xi^2}+\frac{\partial^2\Theta}{\partial\eta^2} \qquad (9)$$

$\bar{u}$ is the cross-sectional average velocity and can be defined by $\bar{u}=\frac{1}{2h}\int_{-h}^{h}u(y)dy$

where, $Pe=\bar{u}h/D$, and $\xi=X-\bar{u}\tau$ are Peclet number and the non-dimension coordinates and parameters moving with the mean velocity $\bar{u}$. The boundary and initial condition (7) takes the form

$$\Theta(0, X, \eta) = 1 \ \text{for} \ |X| \leq \tfrac{1}{2}X_s$$
$$\Theta(0, X, \eta) = 0 \ \text{for} \ |X| \geq \tfrac{1}{2}X_s \qquad (10)$$
$$\Theta(\tau, X, \eta) = 0, \ \frac{\partial\Theta}{\partial\eta}(\tau, X, \eta) = 0 \ at \quad \eta = \pm 1$$

The solution of Eqn. (9) subject to Eqn. (10) is written as a series expansion in $\frac{\partial^k\Theta}{\partial\xi^k}$, in the form

$$\Theta = \Theta_m(\tau, \xi) + \textstyle\sum_{k=1}^{\infty} f_k(\tau, \eta) \frac{\partial^k\Theta_m}{\partial\xi^k} \qquad (11)$$

where, $\Theta_m(\tau, \xi) = \frac{1}{2}\int_{-1}^{1} \Theta \mathrm{d}\eta$ (12)

Substituting (11) in Eqn. (9), we have

$$\frac{\partial\Theta_m}{\partial\tau} + U^* \frac{\partial\Theta_m}{\partial\xi} - \frac{1}{{P_e}^2}\frac{\partial^2\Theta_m}{\partial\xi^2} + \textstyle\sum_{k=1}^{\infty}\left[\left(\frac{\partial f_k}{\partial\tau}\right) - \left(\frac{\partial^2 f_k}{\partial\eta^2}\right)\right]\frac{\partial^k\Theta_m}{\partial\xi^k} + U^*\frac{\partial^{k+1}\Theta_m}{\partial\xi^{k+1}} - f_k Pe^{-2}\frac{\partial^{k+2}\Theta_m}{\partial\xi^{k+2}} + f_k\frac{\partial^{k+1}\Theta_m}{\partial\tau\partial\xi^k} = 0 \qquad (13)$$

where,

$$\frac{\partial\Theta_m}{\partial\tau} = \sum_{k=1}^{\infty} k_k(\tau)\frac{\partial^k\Theta_m}{\partial\xi^k}$$

where the dispersion coefficients $k_k(\tau)$ are time dependent. This form, unlike in the model of Taylor (1953) and Aris (1956) is established on the premise that the cognitive operation of distributing (is diffusive in nature right from time zero. Now Eqn. (13) is solved subject to boundary conditions:

$$\Theta_m(0, \xi) = 1 \ \text{for} \ |\xi| \leq \tfrac{1}{2}X_s \qquad (14)$$
$$\Theta_m(0, \xi) = 0 \ \text{for} \ |\xi| \geq \tfrac{1}{2}X_s \qquad (15)$$
$$\Theta_m(\tau, \xi) = 0 \qquad (16)$$

Using Eqn. (14) into (13), and rearranging terms, we get

$$\left[\frac{\partial f_1}{\partial\tau} - \frac{\partial^2 f_1}{\partial\eta^2} + U^* + k_1(\tau)\right]\frac{\partial\Theta_m}{\partial\xi} + \left[\frac{\partial f_2}{\partial\tau} - \frac{\partial^2 f_2}{\partial\eta^2} + U^* f_1 + f_1 k_1(\tau) + k_2(\tau) - Pe^{-2}\right]\frac{\partial^2\Theta_m}{\partial\xi^2} + \textstyle\sum_{k=1}^{\infty}\left[\frac{\partial f_{k+2}}{\partial\tau} - \frac{\partial^2 f_{k+2}}{\partial\eta^2} + U^* + f_{k+1}k_1(\tau) + (k_2(\tau) - Pe^{-2}) + f_k \sum_{i=3}^{k+2} k_i(\tau) f_{k+2-i}\right]\frac{\partial^{k+2}\Theta_m}{\partial\xi^{k+2}} = 0 \qquad (17)$$

We get,

$$f_1 = \frac{\eta^4}{4} - \eta^3\beta + \frac{\eta^2}{2} - \frac{13}{60} + \textstyle\sum_{n=1}^{\infty} A_n e^{-{\lambda_n}^2\tau}\cos(\lambda_n\eta) \qquad (18)$$

where,

$\lambda_n = n\pi$, and

$$A_n = -{\lambda_n}^{-2}\left[(-1)^n(3\beta - 4) - 6\beta{\lambda_n}^{-2}\right] \qquad (19)$$

On Solving the above equations dispersion coefficient, we get

$$k_2(\tau) = Pe^{-2} - \frac{12\beta^2}{5} - \frac{86}{105} + 12\textstyle\sum_{n=1}^{\infty}\frac{A_n e^{-{\lambda_n}^2\tau}}{{\lambda_n}^2} \qquad (20)$$

Similarly $K_3(\tau)$, $K_4(\tau)$ and so on are obtained. The expressions are omitted here because of their lengthy expression. We find that $k_i(\tau), (i > 2)$ are negligibly small compared to $k_2(\tau)$, the generalized dispersion model reduces to

$$\frac{\partial\Theta_m}{\partial\tau} = K_2(\tau)\frac{\partial^2\Theta_m}{\partial\xi^2}$$

and the solution of above Eqn. is

$$\Theta_m = \frac{1}{2}\left\{erf\left(\frac{\frac{1}{2}X_s - \xi}{2\sqrt{T}}\right) + erf\left(\frac{\frac{1}{2}X_s + \xi}{2\sqrt{T}}\right)\right\}$$

where, T = $\int_0^{\tau} K_2(z)\, dz$

$$T = \left(Pe^{-2} - \frac{12\beta^2}{5} - \frac{86}{105}\right)\tau + 12\sum_{n=1}^{\infty}\frac{A_n(1 - e^{-{\lambda_n}^2\tau})}{{\lambda_n}^4}$$

III. Results and Discussions

The problem of generalized dispersion has been analyzed for a simplified computational model of human knee joints with representation of synovial fluid by the viscous fluid. Fig. 3 is plotted between Taylor's dispersion coefficients $k_2(\tau)$ and instantaneous time τ for different values of viscosity μ of the synovial fluid. It has been depicted from the figure that at the increase value of time the second dispersion coefficient $k_2(\tau)$ also increases. It has been depicted from the fig 3. that the dispersion coefficient increasing with decreasing values of the viscosity μ decreases. The similar result has been obtained by [16] in a dispersion of solutes in a in the laminar flow between two parallel plates by taking into consideration the homogeneous and heterogeneous reaction of the solvent with the solute. In the case of a diabetic patient as compared to normal subject the viscosity of the plasma is higher; in turn the diffusion coefficient for the diabetic person is generally higher. Fig. 4 shows the variation of mean concentration distribution Θ_m with axial distance for different values of viscosity μ.

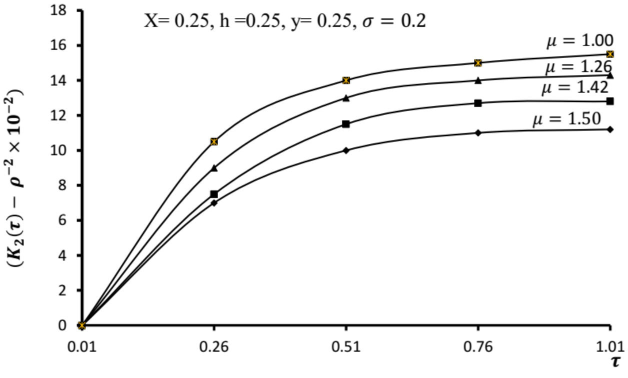


Fig. 3. Variation of dispersion coefficient with time τ for different values of the of viscosity

The Fig. 4 shows clearly that mean concentration Θ_m decreases asymptotically as axial distance approaches to infinity. It should be noted that as viscosity μ increases then mean concentration also increases [17, 21]. The concentration of hyaluronic acid molecules increase due to increase value of the viscosity of the synovial fluid, eventually it increases the apparent viscosity of the lubricant i.e. synovial fluid. Fig. 5 depict the variation between the non-dimensional concentration distribution of solute with non-dimensional time for various values of viscosity. It is clear from the figure that the mean concentration distribution Θ_m decreases with increases values of the time τ [21, 22].

In synovial cavity, enhanced diffusion may occur in the initial phases of the movement as the solute only moves through the larger interstitial spaces. As diffusion progresses the solute may move into the smaller interstitial volumes.

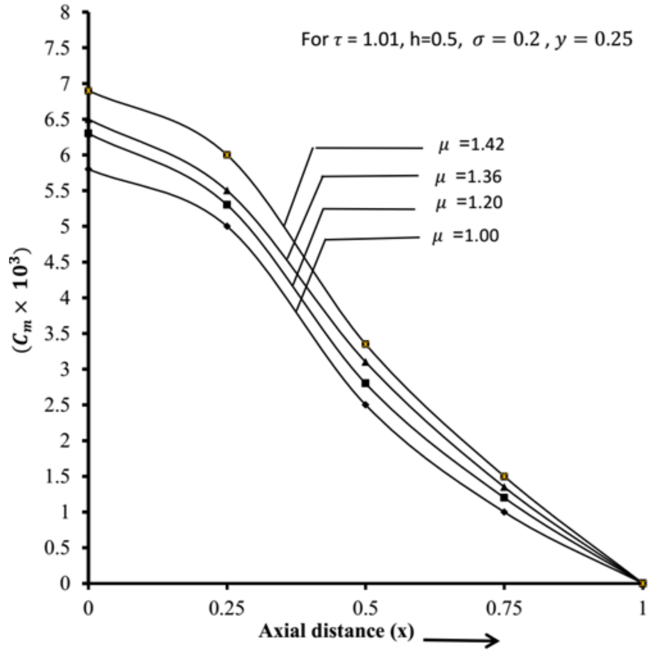


Fig. 4. Variation of mean concentration distribution with axial distance for different values of viscosity (μ)

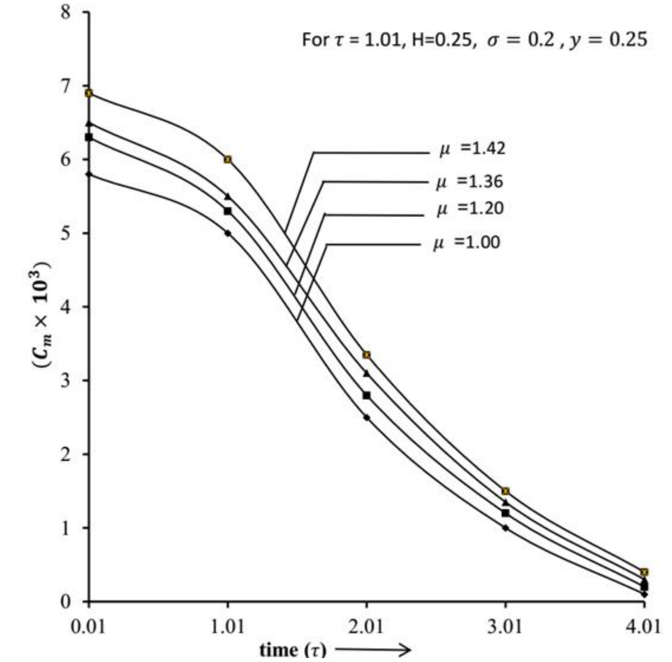


Fig. 5. Variation of mean concentration with time (τ) for different values of viscosity (μ)

IV. Conclusions

The dispersion of proteins and other nutrients from the synovial fluid to articular cartilage is studied using the [18] exact analysis of unsteady convective diffusion. It has also been observed that dispersion coefficient $k_2(\tau)$ increases with a decrease in the viscosity. It is seen that the mean concentration distribution decreases Θ_m with an increase in the axial distance and in the time and increases with increase in viscosity.

It is seen that the mean concentration distribution decreases with an increase in the time and axial distance the cells of middle area get more nutritional as compared to the peripheral area. It helps to orthopedic surgeons to check by the formula of dispersion mechanism, whether the joints functioning effectively or not. In the future, the model for unsteady convective diffusion can be used for the development of a mathematical model for the articular cartilage regeneration because the key mechanism involved in the cartilage regeneration modeling cell migration, nutrient diffusion and depletion extracellular matrix synthesis and degradation at the defect site, both spatially and temporally.

References

[1] Mow V.C. and Ateshian G. A. (1997) ‘Lubrication and wear of Diarthrodial joints’ Basic Orthopaedic Biomechanics, V.C. Mow and Hayes W.C. ed., Lippincott-Raven, Philadelphia, pp. 273-315.

[2] Bali R. and Shukla A.K. (2000), ‘Rheological effects of synovial fluid on nutritional transport’ Tribology Letters Vol. 9, No. 3-4, pp. 233-239.

[3] Jin Z.M. Dawson. D. and Fisher J. (1992) ‘The effect of porosity of articular cartilage on the lubrication of a normal human hip joint’ Proceeding of Institution of Mechanical Engineers II: Journal of Engineering and Medicine, Vol. 206, pp. 117-124.

[4] Loret B. and Fernando. M.F. (2005) ‘A framework for deformation, generalized diffusion, mass transfer and growth factor in multispecies multiphase, biological tissues’ European journal of mechanics and Solid, Vol. 24, pp. 757-781.

[5] Shiroky A., Alexander V., Volkov V., Novochadov V., (2014) ‘Crucial Processes Interaction During the Renewal of Articular Cartilage: the Mathematical Modeling’ Alexander European Journal of Molecular Biotechnology, Vol. 4, No. 2, pp. 86-94.

[6] Jaiswal D. K., Kumar A, Kumar N, Singh M. K. (2011) ‘Solute transport along temporally and spatiall dependent flows through horizontal semi-infinite media: dispersion proportional to square of velocity’ ASCE J Hydrol Eng 16(3):228–238

[7] Kim S, Kavvas ML (2006) ‘Generalized fick’s law and fractional ade of pollution transport in a river: detailed derivation’ ASCE J Hydrol Eng 11(1):80–83

[8] Videcoq P, Steenkeste K, Bonnina E, Garnier C (2013) A multi-scale study of enzyme diffusion in macromolecular solutions and physical gels of pectin polysaccharides. Soft Matter 9:5110–5118

[9] Gupta BP, Thakur N, Jain NP, Banweer J, Jain S (2010) Osmotically controlled drug delivery system with associated drugs. J Pharm Pharm Sci 13(3):571–588

[10] Gilbert Makanda, Sachin Shaw, and Precious Sibanda (2015) ‘Diffusion of Chemically Reactive Species in Casson Fluid Flow over an Unsteady Stretching Surface in Porous Medium in the Presence of a Magnetic Field’ Mathematical Problems in Engineering, Article ID 724596

[11] Tripathi D, Yadav A, Anwar O.B . (2017) Electro-kinetically driven peristaltic transport of viscoelastic physiological fluids through a finite length capillary: Mathematical modeling. Mathematical Biosciences 283, 155-168.

[12] Brault A, Dumas L, Lucor D (2016) ‘Uncertainty quantification of inflow boundary condition and proximal arterial stiffness coupled effect on pulse wave propagation in a vascular network’ arXiv:1606.06556

[13] Kohlhof H et al (2016) ‘Single Molecule Microscopy Reveals an Increased Hyaluronan Diffusion Rate in Synovial Fluid from Knees Affected by Osteoarthritis’ Nature Scientific Reports DOI: 10.1038/srep21616.

[14] Walker P.S. Dowson D. Longfield M.D. and Wright V. (1968) ‘Boosted lubrication in synovial joints by fluid entrapment and enrichment’ Annuls of the Rheumatic Diseases, Vol. 27, pp. 512-520.

[15] Ward A.C. Dowthwalte G.P. and Pitsllides A.A. (1999) ‘Hyduron in joint cavitation’ Biomech. Soc. Trans., Vol. 27, pp.128-135.

[16] Chandra P. and Agarwal R.P. (1983) ‘Dispersion in simple microfluid flows’, Int. J. Engng. Sci., Vol. 21 No. 5, pp.431-441.

[17] Bailo P, Van A. and Meulen, M (2001) ‘A mathematical framework to study the effects of growth factor influences on fracture healing’ Journal of Theoretical Biology, Vol. 212 (2), 191–209.

[18] Taylor G.I. (1953) ‘Dispersion of soluble matter in solvent flowing slowly through tube’ Proc, Roy, Soc. Lond., A, Vol. 219, pp.186-203.

[19] Aris R. (1956) ‘On the dispersion of a solute in fluid flow through a tube’ Proc. Soc. London A Vol. (235), pp.67-77.

[20] Sueiu, A.N., Iwatsubo, T., and Matsuda, M., (2003) ‘Theoretical investigation of an artificial joint with micro-pocket-covered component and biphatic cartilage on the opposite articulate surfac’ ASME J. Biomech. Eng., Vol. 125, pp. 425-433.

[21] Lutianov M. Naire S. Roberts S. Kuiper J. (2011) ‘A mathematical model of cartilage regeneration after cell therapy’ Journal of Theoretical Biology 289 pp.136–150.

[22] Zhou, S., Cui, Z., Urban, J., (2007) ‘Nutrient gradients in engineered cartilage: metabolic kinetics measurement and mass transfer modeling’ Arthritis and Rheumatism Vol. 50 (12), 3915–3924.

[23] Rudraiah N., Raghunatha S.V., (2013) ‘Dispersion in Chiral Fluid in the Presence of Convective Current between Two Parallel Plates Bounded by Rigid Permeable Walls’ Journal of Applied Fluid Mechanics, Vol. 6, No. 1, pp. 7-13.

[24] Rudraiah1 N, Mallika K. S. and Sujatha N. (2016) ‘Electrohydrodynamic Dispersion with Interphase Mass Transfer in a Poorly Conducting Couple Stress Fluid Bounded by Porous Layers’ J. of Applied Fluid Mechanics, Vol. 9, No. 1, pp. 71-81.

[25] Prathap J. K, J.C. Umavathi and Madhavara S. (2012) ‘Effect of homogeneous and heterogeneous reactions on the solute dispersion in composite porous medium’ International Journal of Engineering, Science and Technology Vol. 4, No. 2, , pp. 58-76.

[26] Gill W.N. and SankaraSubramanian R. (1970) ‘Exact analysis of unsteady convective diffusion’ Proc. Roy. Soc. Lond., Vol. 316A, pp.341-350.

Multi Objective Optimization of Cloud Computing Services for Consumers

Eli WEINTRAUB
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

Yuval COHEN
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

Abstract—**This paper presents a novel multi objective model for optimizing the purchase decision of a cloud computing services customer. The providers are typically offering consumers cloud computing varying information systems services. The cloud services consist of different functionalities at varying costs, and varying reliability. So the customer's main objectives (based on the literature) are to maximize their utility, and minimize their costs and risks. Since utility cost and risks are different dimensions, the problem is essentially a multi-objective optimization problem. So far, previous research does not address the multi objective nature of the problem. This article deals with optimizing consumers' decision, but at the same time maintaining each of their objectives' considerations. An optimization model presented and illustrated. The article also demonstrates the advantages gained by the optimization model when implemented using the dynamic cloud architecture over the traditional cloud architecture.**

Keywords—Cloud Computing; Security Risk; Software as a service; Platform as a service; Infrastructure as a service; Optimization; Cost; Utility

I. INTRODUCTION

There are several definitions for Cloud Computing (CC). In this work we use NISTs' definition [1] as an on-demand convenient remote access to a pool of computing resources managed by a CC service provider.

In the past, organizations managed their computing resources inside their geographical borders. In the last years more organization move their servers outside their firms' borders, to Service Providers (SP) who take responsibility of various computing activities managing the computing resources and facilitating the services. Doing so, organizations are facing new risks and problems which they have not met in the past. Reference [2] claims that organizations have to make changes in the production processes, defining new risk management procedures, and changing their IT management processes.

CC services are being used by four kinds of organizations: public, community, hybrid and private [3]. Public organizations locate their computing resources inside their geographic borders or outside it, at the cloud services providers' site, after considering issues of privacy, security, ease of production and financial. Community services are aimed at a group of consumers who have similar interests, buying resources from one external service provider. Hybrid services enable consumers using their internal resources in parallel to outside providers. Private customers, mostly locate resources at the providers' site. CC providers allocate their resources, which they supply to various consumers, trying to make a total safe separation of data and processes belonging to different organizations.

This article reviews the main advantages of using CC model and reviews the barriers and risks adopting the CC model. The information security issue is mentioned in literature as a barrier to CC adoption, and is an issue dealt largely in CC research [4].

Consumers' buying decisions of CC services are not simple. Providers are typically offering consumers cloud computing varying services, difficult for comparison. The services consist of different functionalities at varying costs, and varying reliability. So the customer's main objectives are maximization of their utility, and minimization their costs and risks. Since utility, cost and risks are different dimensions, the problem is essentially a multi-objective optimization problem. Published research does not address the multi objective nature of the problem. This article deals with optimizing consumers' decision, but at the same time maintaining each of their objectives' considerations. The article reviews the known models dealing with a single objective optimization decision. In this article, a novel multi-objective optimization model presented and illustrated. The article also demonstrates the advantages gained by the optimization model when implemented using the dynamic cloud architecture.

The article is organized as follows: Section II is an overview of the current CC architecture and the dynamic network architecture, which is used by the model. Section III reviews consumers' buying considerations. Section IV is an overview of cost optimization, Section V reviews utility optimization. Section VI reviews risk optimization. Section VII presents and illustrates the Multi Objective Optimization model proposed in this paper. Finally, section VIII concludes and suggests future possible research directions.

II. CLOUD COMPUTING ARCHITECTURE

CC architecture consists of three layers: Infrastructure (IaaS), Platform (PaaS) and Software application (SaaS). Each layer is responsible for delivery of certain services for consumers. Each layer also fulfils the requests of the upper layers. A framework of the CC architecture is defined by [5], composed of three layers in parallel to functions supporting CC services. Figure I describes current CC architecture.

Rectangles describe computing services. The organization buys all CC services from one SP.

The functions of each layer are as follows:

Infrastructure layer – This layer provides basic technologies as hardware, communication resources, operating systems and systems' utilities.

Platform layer – This layer operates on the top of the infrastructure layer, providing platform services such as development environments and business platforms.

Application layer – This layer operates on top of the platform and infrastructure layers, providing applications software and human interfaces used by the organizations' end-users and customers.

Service providers offer their services in bundles. A consumer buying a SaaS service will have to use the PaaS and IaaS services offered by the SP. A consumer wishing to buy a PaaS service will have to use also the providers' IaaS services. The bundling practice forbids consumers who wish to consume certain services from different SPs. According to [5], nowadays, certain providers use to run applications running on other providers' infrastructure, but the consumer is blind to this separation of platforms while buying his service from one single SP. The bundling practice limits free market forces from competing in this king of services, forcing customers pay for services they may buy from other providers in cheaper prices. For example, a consumer may buy a PaaS service from SP1, but the underlying IaaS service from a SP2, which sells the appropriate infrastructure service cheaper than SP1. According to [6], in the future, application will be designed including modularity which will enable running parts of the application on different SPs' platforms. Ref. [5] states that the cloud computing architecture is more modular compared to traditional hosting architectures, which might be a byproduct of the CC three layers' architecture. CC components are loosely coupled, thus enabling the development of modular applications which enable distribution the application among several SPs. Ref. [7] also claims that applications belonging to different layers will run on separate geographical locations. Ref. [8] claims that virtual machine hardware allows transfer of applications to other machines, provided by different IaaS providers. Ref. [9] suggests to make use of multiple clouds, achieving security targets.

This article continues the research direction proposed in [10] basing CC services on a dynamic business model. According to the dynamic model a consumer is able to buy certain SaaS services using SP1 resources and buy PaaS or IaaS services from other service providers. Implementation of the dynamic architecture needs technological standardization of the interfaces, which enables improved interoperability and connectivity of applications' components. Also, systems' building blocks should implement loose coupling principles. Following those design principles will enable connectivity among vertical and horizontal services, thus eliminating the bundling phenomena. References [10] [11] demonstrate the advantages achieved in aspects of consumers' cost and utility optimization, based on the dynamic architectural model. Figure II presents the dynamic CC architecture, describing consumers served by different providers for each layer and service. Arrows describe services supplied by underlying layers. Rectangles describe CC services. The business consumes its CC services from many SP's choosing the best combination of service providers.

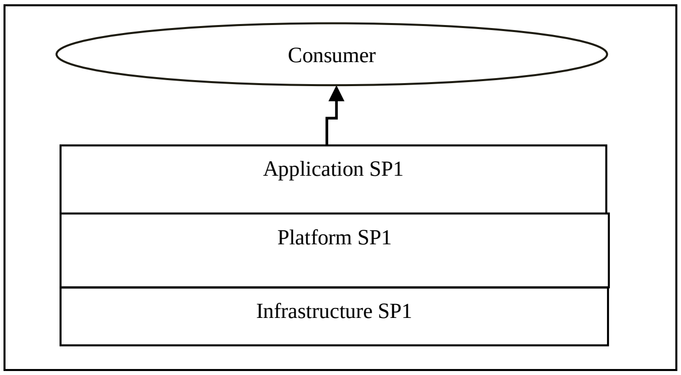


Fig. 1. Cloud Computing Current Architecture – one service provider

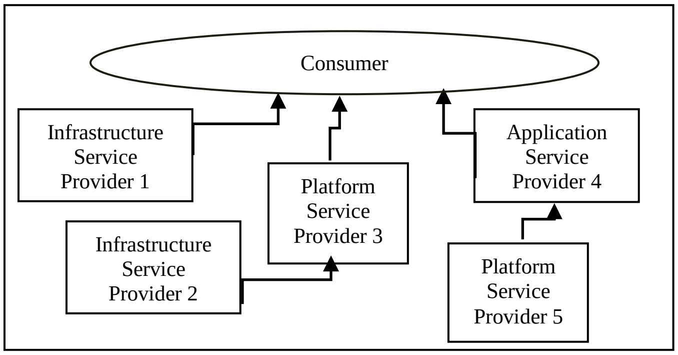

Consumer
Infrastructure Service Provider 1
Platform Service Provider 3
Application Service Provider 4
Infrastructure Service Provider 2
Platform Service Provider 5

Fig. 2. Cloud Computing Dynamic Architecture

III. Buying Considerations

Organizations use varying criteria for their CC buying decisions. There are organizations emphasizing costs, other emphasize risks while others consider the overall utility in their CC adoption decisions. Ref. [12] found that financial organizations regard CC a cost-effective technology which contributes to their capital efficiency. The researchers also found that financial organizations regard security as a barrier to CC adoption, among other risk factors.

Comparisons of pricing models of CC services is an issue researched largely, but variations among the structures of the pricing schemes puts major difficulties in coming to clear conclusions [14]. There are different viewpoints on the issue of CC costs. Ref. [15] found that organizations regard cost savings as the first adoption motivation, but the least researched issue, although research interests are rising.

Lack transparency of the resources supplied by service providers are regarded a key risk factor for organizations considering CC adoption. Several researchers studied the transparency issue. In one research, public cloud consumers got no permissions to view IT infrastructures, in other cases, consumers got partial views of resource consumption [16] [17]. Consumers wishing to make predictions concerning their future CC cost have difficulties because of the transparency issue, and lack of monitoring tools. According to [18] there is little research on the monitoring and prediction strategies in the CC domain. Researchers suggest handling the transparency issue by introducing pricing models presenting all components' prices of each service [19]. Other researchers state that currently, varying pricing models and large numbers of CC providers lead to complexities in the adoption decision process [20].

Cost and risk minimization in buying decisions lacks considerations of the vast advantages of CC model. Several organizations focus on CC utilities or a mix of expenses and utility considerations. Several organizations prefer to compare the utilities rather than expenses of risk adoption. In those cases, consumers face the same kind of difficulties stemming from the issue of insufficient transparency. Utility criteria selection might be complicated to measure and compare since providers offer different services having various functionalities, on un-standard scales. Various techniques have been suggested simulating consumer utility decisions. There are several techniques coping this purpose. Literature describes conjoint analysis a useful methodology, which enables coping with providers' selection issue.

To conclude, there is much research showing a large variance in usage of criteria lists used for CC adoption decisions. CC decisions involves complicated decisions with no standard scales assisting consumers in performing evaluations assisting management decisions. We categorize the decision factors to three main kinds: cost minimization, risk minimization, and utility maximization, each category consists of specific characteristics.

IV. Cost Optimization

Literature describes two principal pricing models. The pay-per-use model is the popular model, and the second fixed-price model [21]. In the pay-per-use model consumers pay a fee according to the price of the resource, duration and volume consumed. Resources are IT components such as hardware, operating system, database, e-mail or enterprise application [22]. Volumes are specified as resource units such as processor seconds, disk or memory gigabytes, number of printed pages etc'. In the pay-per-use model consumers are not limited to the volume or duration they use, although some agreements limit volumes to a maximal amount above which the service stops. In the fixed-price model, the consumer pays for the resource consumed irrespective of the duration and volume. This model defines only the period (usually month or year) the consumer may use the service. In case the user does not use all the volume he planned to, he will pay the fee although he had not consumed it. Researchers state that the pay-per-use model is a better driver to free market competition and to efficient computing resources allocation in the CC market [23]. Researchers state that the pay-per-use model is the current CC market trend direction [24].

Researchers found several kinds of anomalies during consumers' decisions concerning choosing a costing model. Some consumers prefer to pay more for a fixed-price scheme for volumes they may not use [25]. Ref. [26] defined a pricing model called a bursting model, which balances consumers' varying demands to computing resources by switching resource workloads among consumers, offering consumers a stable quality of service. This model assists consumers solving dilemmas caused by reaching the maximal package capacity in cases of high workloads.

Researchers found biases of two kinds: irrational economic decisions preferring the pay-per-use model and the opposite irrational decision. Ref. [27] found that consumers are paying more in a fixed-price model for budget planning argumentations, and found cases of consumers choosing the pay-per-use model and actually paying more, for reasons of inability to predict future resource consumption. Ref [28] who studied consumers' pricing models also states that the fixed-price biased decisions, were influenced by budgetary argumentations, while the pay-per-use biased decisions were influenced by the productions' flexibility motivation.

Pricing biases stemming from providers' interests are described in research literature. Providers are mainly interested in marketing and strategic reasons. Providers differentiate users' pricing schemes, offering cheap prices, sometimes free of charge, or high service level agreements to consumers they are interested to attract or lock-in [28] [29]. Providers are customizing special software features to certain customers, which need also programming changes in consumers' software, thus causing them high switching costs when they are considering leaving to other providers [30].

Ref. [10] proposes a framework, which enables to compare different tariff tables of different SPs on one unified scale, thus optimizing costs. Nevertheless, that framework does not deal with other buying considerations such as risk and utility.

V. Utility Optimization

There are several techniques enabling utility comparison of different services and products, such as Cluster Analysis and Multidimensional Scaling (MDS) [31]. Conjoint analysis is a methodology which enables to analyze buying trade-offs considerations among competing products [32]. The methodology makes use of a technique examining the characteristics of each product, simulating and predicting buyers' considerations while comparing different products. A study which used conjoint analysis methodology described in [33] found that the most influencing CC buying decisions' characteristics were quality of service and lock-in prevention. Researchers found that information security is a factor in CC adoption considerations [34]. Ref. [35] states that consumers are shifting from technological to service-oriented issues in their CC adoption considerations. In a survey [36], researchers found that the consumers mentioned six attribute levels: (1) providers' reputation, (2) required skills, (3) migration process, (4) pricing tariff, (5) cost compared to internal solution and (6) consumer support. Security is an adopting barrier to CC services [37]. Ref. [36] who used conjoint analysis, found that providers' reputation was the attribute with the highest relative importance of 26%, migration process was the second with 21% importance. Cost has been found fourth having 16% relative importance. Researchers who studied service attributes influencing on CC adoption, found seven groups of attributes: Monetary payoff, usability, flexibility, trademark, added value, connectivity and customers' support [38].

To conclude, there has not been found one single list of utility attributes, nor one agreed methodology for utility comparisons. Ref. [39] [11] describe a methodology which enables utility attributes comparison for consumers in the CC dynamic environment, but lacking cost and risk considerations.

VI. Risk Optimization

Risk assessment in the CC domain is an issue dealt intensively in literature [40] [41]. Researchers state that security risks are among the biggest obstacles to adoption cloud services [9].

This article focuses on security risks since this risk category is a major inhibitor of CC adoption, without limiting generality of the proposed model. Security risks as a subgroup of the outsourcing issue are a complex research area, which researchers still are not able to fully capture it's complex nature [42]. Ref. [43] States that managing security risks is getting more complex, and many publications include proposals targeting the various cloud security threats.

Cloud security covers several categories. Ref. [44] surveyed the research publications on cloud security issues, identified the basic concepts underlying vulnerabilities and threats, and classified them as follows: virtualization elements, multi-tenancy, cloud platform and software, data outsourcing, data storage security and standardization and trust. Ref. [45] also categorized security risks to three kinds: Multiple Users, Minimal Control and Single Point of control. Ref. [43] presents a method to assess security risks and a set of steps to identify and assess security risks. Accordingly, risks are categorized to Six-View Perspectives: Threat view, Resource View, Process View, Risk Assessment View, Management View, and Legal View. Ref. [39] presents a security risk assessment model based on ISACA's framework defined [46]. The framework is designed to present a practical guidance for IT and business professionals concerning the decision to move to the cloud. The guide provides checklists outlining the security factors considered when evaluating the cloud as a potential solution.

Ref. [39] proposes a model, assisting consumer in assessing risks, but does not handle decision factors of cost and utility maximization.

To summarize, there is no one integrative model enabling decision support for managements who wish to compare and evaluate all CC attributes, naming costs, risks and utility.

Next, we present our Multi Objective Optimization model.

VII. The Multi Objective Optimization Model

This section analyzes the CC consumer choice as a multi-dimensional model and propose a structured approach to eliminate alternatives and choose the best option.

There are three main objectives that a CC user seeks to optimize when choosing service provider:

1) Maximal utility
2) Minimal cost
3) Minimal risk

The different dimensions of these objectives are an obstacle in the way to form a simple model for decision making (mainly choice of service provider and a bundle of services). While translating everything to money is possible – it is typically subjective and far from accurate. The same could be said about a fitness function.

We start optimization computations using the original data of each dimension (naming cost, utility and risk) which was computed according to its specific characteristics. Cost computed according [10], utility computed according to [11], and risk scores computed according to [39].

We now present the multi objective optimization model in two business models. First the case of bundled services of all three layers in which an organization buys all CC services from one single service provider, implemented on the current CC architecture as described in Fig. I. Second, the business model of a free market – Choosing the best provider per layer, among all SPs. This business model is implemented on the dynamic CC architecture as described in Fig. II.

- **The case of bundled services of all three layers**

CC has three fundamental layers (IaaS, PaaS, SaaS), and the SPs traditionally try to bundle their services in all the layers so as to bound the consumer to them in all the three layers. In this case, the customer needs to consider and choose only one of the various potential SPs. Suppose we have *n* SPs to compare: each one offers a bundle of services that contributes a utility to the consumer, with associated risks, for a given cost. Therefore, comparing *n* SPs is done by comparing their *n* three-dimensional points. It is therefore essential that we develop the mechanism to deal with such three dimensional comparisons.

On the other hand, if we assume the free market forces will enable purchasing services for each layer independently from the other layer, we would have to repeat the choice between n_i providers three times (i=1,2,3): each layer comparison is based on the number of SPs providing that layer– each one is a three dimensional model.

The proposed method is based on rescaling each score to a common scale, which enable a common graphical representation (the original values are retained, and could be used if necessary). This approach retains three separate dimensions for the comparison (unlike translating the objective to one fitness function, or to monetary value).

The method works as follows: for each dimension, we rescale the best performance to be 10, and the least performance to be 5. The other scores are then interpolated in that range. We then rule out suppliers having worst performance in any dimension. This process of ruling out or eliminating SPs continues until the last one is left.

The following example illustrates the suggested techniques. Table 1 includes the original values of utility, cost and risk score for five SPs. Tables 2,3 normalize the original values to one common scale. Fig. III presents graphically the normalized objective values.

TABLE I. ORIGINAL DATA

Supplier	Annual Utility	Annual Cost	Annual Risk
SP1	8	$ 50,000	12
SP2	7	$ 60,000	8
SP3	10	$ 70,000	15
SP4	9	$ 50,000	10
SP5	8	$ 40,000	9

TABLE II. COMMON SCALE COMPUTATIONS

Supplier:	Annual Utility	Annual Cost	Annual Risk
Optimal	10	40,000	8
Least Optimal	7	70,000	15
Range	3	30,000	7
New max	10	10	10
New min	5	5	5
New range scale	3/5	30,000/5=6,000	7/5

TABLE III. COMMON SCALE

Supplier	Annual Utility	Annual Cost	Annual Risk
SP1	6.67	8.33	7.14
SP2	5.00	6.67	10.00
SP3	10.00	5.00	5.00
SP4	8.33	8.33	8.57
SP5	6.67	10.00	9.29

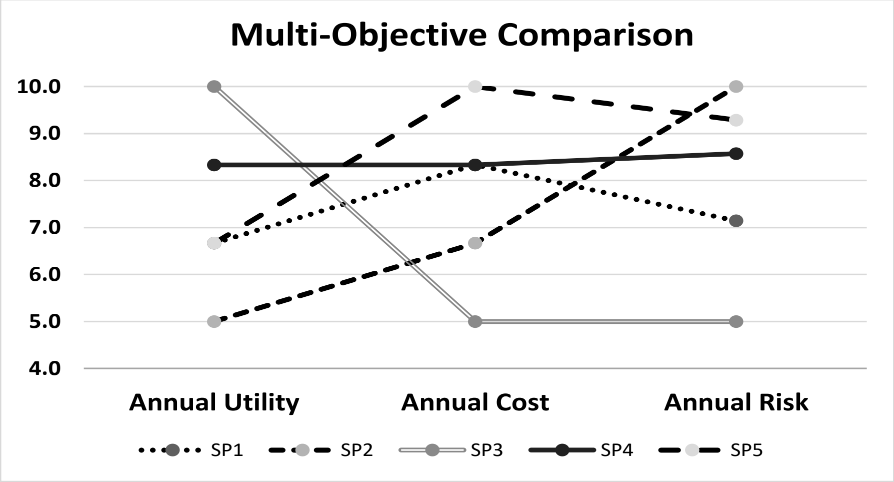


Fig. 3. Graphical representation of the objective dimensions of 5 SPs

It is clear from Fig. III above that:

In this case SP3 is worse in terms of Cost and Risk, and therefore is eliminated.

SP2 is worse in terms of Utility, and therefore is eliminated.

The remaining suppliers (1, 4 and 5) are depicted in Fig. IV.

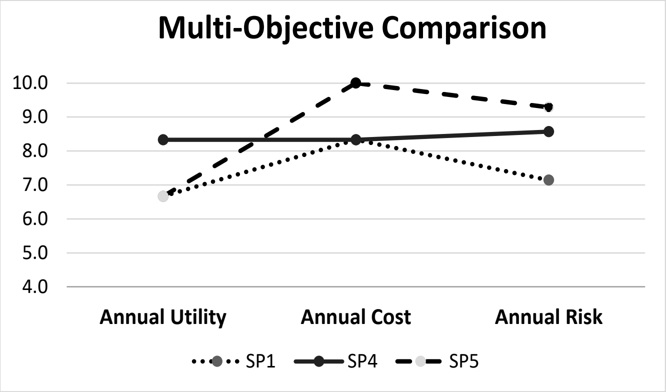


Fig. 4. Graphical representation of the objective dimensions of remaining SPs

From Fig. IV it is clear that SP1 form the lower envelop and is eliminated. Then for the remaining SP4 and SP5: SP4 has two minimal points while SP5 only one, so SP5 remains the best option.

Another approach would be using the original numbers of each remaining supplier (as in Table 4).

TABLE IV. THE LAST REMAINING SPS

Supplier	Annual Utility	Annual Cost	Annual Risk
SP4	9	$ 50,000	10
SP5	8	$ 40,000	9

So the final tradeoff is between a unit of utility vs. $10,000 plus a unit of risk.

Here a subjective decision could be taken, based on the preference of the individual.

Another possibility is to design simple decision rule to choose between SP4 and SP5. For example, if each dimension has the same importance, the scales in table 2 could be used:

3/5 Utility = $6,000 = 7/5 Risk. So 1 Utility = $ 10,000 = 7/3 Risk, and SP5 is chosen over SP4 since: 1 Utility < $ 10,000 +1 Risk

- **The case of a free market –Choosing provider per layer**

The practice of SPs to bundle their services in all three layers into one offering dictate a choice and a contract with a single SP. In contrast to the bundled services, free market forces should enable customers acquire services in each layer, independently from the other layers. In the long run we would have to repeat the choice between n_i providers for each layer – that is three times (i=1,2,3): each layer comparison is based on the number of SPs providing that layer– each combination of a layer and SP is still a three dimensional point. Table 5 describes the break-down of Table 1 into the three layers.

TABLE V. ORIGINAL DATA OF TABLE 1 DETAILED BY LAYER

Supplier	Annual Utility			Annual Cost			Annual Risk		
	IaaS	PaaS	SaaS	IaaS	PaaS	SaaS	IaaS	PaaS	SaaS
SP1	2	3	3	$ 15,000	$ 15,000	$ 20,000	4	4	4
SP2	2	2	3	$ 15,000	$ 20,000	$ 25,000	2	3	3
SP3	3	3	4	$ 15,000	$ 25,000	$ 30,000	5	5	5
SP4	3	3	3	$ 20,000	$ 15,000	$ 15,000	3	4	4
SP5	3	3	2	$ 10,000	$ 15,000	$ 15,000	3	3	3

Table 5 is conveniently broken down into three layers as described in tables 6, 7, and 8. Tables 6,7 present the detailed computations of IaaS optimization. Fig. V presents IaaS SPs comparison on a Common Scale.

Finding the best IaaS SP

TABLE VI. ORIGINAL **IAAS** DATA BY LAYER (FROM TABLE 5)

Supplier	Annual Utility	Annual Cost	Annual Risk
	IaaS	IaaS	IaaS
SP1	2	$ 15,000	4
SP2	2	$ 15,000	2
SP3	3	$ 15,000	5
SP4	3	$ 20,000	3
SP5	3	$ 10,000	3

TABLE VII. **IAAS** COMMON SCALE

Supplier	Annual Utility	Annual Cost	Annual Risk
SP1	5.0	7.5	6.7
SP2	5.0	7.5	10.0
SP3	10.0	7.5	5.0
SP4	10.0	5.0	8.3
SP5	10.0	10.0	8.3

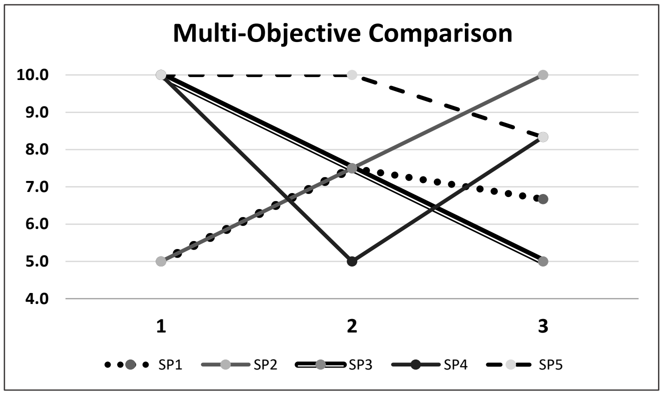


Fig. 5. IaaS SPs Comparison on a Common Scale

It is easy to see in Fig. V above that SP1 and SP2 have minimal utility point, SP3 has maximal risk point and SP4 has maximum cost point. Thus, only SP5 is not eliminated, and is the best IaaS choice.

This procedure repeats twice more for the PaaS and SaaS layers and yields the following. Tables 8, 9, 10, 11 and Fig. VI, Fig. VII follow similar computations for PaaS and SaaS layers.

Finding the best PaaS SP

TABLE VIII. ORIGINAL **PAAS** DATA BY LAYER (FROM TABLE 5)

Supplier	Annual Utility	Annual Cost	Annual Risk
	PaaS	PaaS	PaaS
SP1	3	$ 15,000	4
SP2	2	$ 20,000	3
SP3	3	$ 25,000	5
SP4	3	$ 15,000	4
SP5	3	$ 15,000	3

TABLE IX. **PAAS** COMMON SCALE

Supplier	Annual Utility	Annual Cost	Annual Risk
SP1	10	10	7.5
SP2	5	7.5	10
SP3	10	5	5
SP4	10	10	7.5
SP5	10	10	10

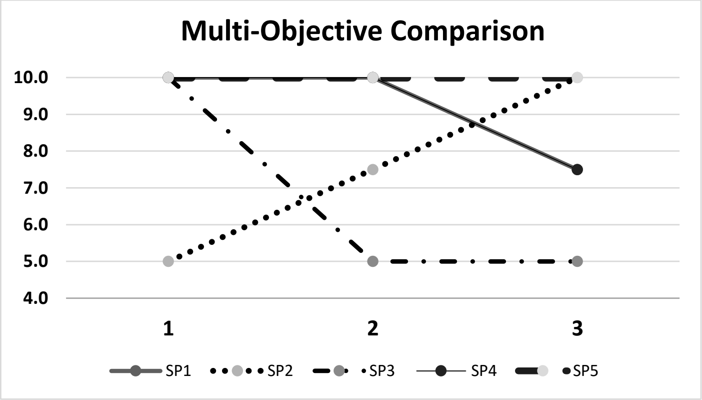


Fig. 6. PaaS SPs Comparison on a Common Scale

In this case, it is easy to see that SP5 dominates all other SPs and is the preferred choice.

Finding the best SaaS SP

TABLE X. ORIGINAL **SAAS** DATA BY LAYER (FROM TABLE 5)

Supplier	Annual Utility	Annual Cost	Annual Risk
	SaaS	SaaS	SaaS
SP1	3	$ 20,000	4
SP2	3	$ 25,000	3
SP3	4	$ 30,000	5
SP4	3	$ 15,000	4
SP5	2	$ 15,000	3

TABLE XI. **SAAS** COMMON SCALE

Supplier	Annual Utility	Annual Cost	Annual Risk
SP1	7.5	8.3	7.5
SP2	7.5	6.7	10.0
SP3	10.0	5.0	5.0
SP4	7.5	10.0	7.5
SP5	5.0	10.0	10.0

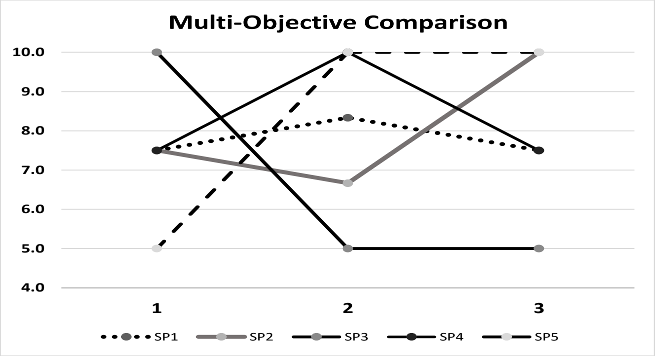


Fig. 7. SaaS SPs Comparison on a Common Scale

It could be inferred from Fig. VII that SP5 should be eliminated due to minimal Utility. SP3 should be eliminated due to maximal risk and cost. Then, SP2 should be eliminated due to minimal remaining value of utility and maximal remaining cost. The remaining alternatives are SP1 and SP4. Since SP4 has lower cost than SP1, and have identical utility and risk. Thus, SP4 is dominating SP1 in SaaS and is the chosen alternative.

Synthesizing the choices at each layer we have: IaaS: SP5; PaaS – SP5; SaaS – SP4. Table 12 summarizes this optimal choice.

TABLE XII. ORIGINAL DATA FOR OPTIMAL CHOICE FOR EACH LAYER: IAAS: SP5; PAAS – SP5; SAAS – SP4

Supplier	Annual Utility			Annual Cost			Annual Risk		
	IaaS	PaaS	SaaS	IaaS	PaaS	SaaS	IaaS	PaaS	SaaS
SP4			**3**			**$ 15,000**			**4**
SP5	**3**	**3**		**$ 10,000**	**$ 15,000**		**3**	**3**	
Total	**3**	**3**	**3**	**$ 10,000**	**$ 15,000**	**$ 15,000**	**3**	**3**	**4**
Total Utility			**9**	**Total Cost**		**$ 40,000**	**Total Risk**		**10**

Comparing to the service bundling case where pure SP5 was chosen we could see that optimizing each of the three layers we get more utility at additional risk, as presented in Table 13.

TABLE XIII. MULTI OBJECTIVE COMPARISON

Supplier	Annual Utility	Annual Cost	Annual Risk
SP5	8<9	$ 40,000 (same)	9<10

In general, optimizing each of the three layers is bound to give either comparable or better results than the choice of a single SP.

VIII. CONCLUSIONS

This paper examines the acquisition of CC services from the perspective of CC customers. While minimizing cost is easy and popular objective for the customers, it captures only a part of the customer's considerations. Maximizing the utility of the customer could be a more inclusive alternative, but since the conversion of money to utility is a tricky business, cost (which could be part of the utility function) is better off as a separate objective. Minimizing risk is a third consideration which is not well suited for conversion to either utility or cost, and thus is a third major objective. Thus, the CC consumer is simultaneously maximizing its utility and minimizing the cost and the risks. Accordingly, this paper presents a multi-objective optimization approach.

While much research has been conducted on CC consumers' decisions (for assessing and optimizing providers' services), current models enable optimizing each dimension separately (cost, utility and risk) on its own scale. Since we did not find in the literature any multi objective models optimizing consumer CC service acquisition. Thus, for the best of our knowledge this is the first time that multi-objective optimization is applied to CC service acquisition.

The proposed model makes use of the dynamic CC architecture, which enables consumers to buy services from several SPs. each one offering services of different layers. We have shown that basing on the dynamic CC architecture organizations can achieve superior advantages relative to the current CC architecture.

Further research is possible in several directions. First, defining a model performing a sensitivity analysis for changes in each dimension. Second, studying ways which assist consumers in assigning their importance weights to their decisions' dimension, which are currently performed subjectively and intuitively. Third, studying ways assisting organizations assess future values of decisions dimensions.

Currently, organizations assess their future CC costs for making their CC adoption decisions (also their utility and risk scores) according to general knowledge, not relying on objective quantitative measures, sometimes irrelevant to their specific current configuration. It is hoped that this paper will contribute to more structured and quantitatively based decisions.

REFERENCES

[1] P. Mell, and T. Grance, “The NIST definition of cloud computing”, National Institute of Standards and Technology, NIST, Vol. 53 No. 6, p. 50, 2009.

[2] T. Pueschel, A. Anandasivam, S. Buschek, and D. Neumann, "Making money with clouds: Revenue optimization through automated policy decisions". ECIS - European Conference on Information Systems 17, 2009.

[3] C. Weinhardt, B. Blau, and J. Stößer, "Cloud Computing – A Classification, Business Models, and Research Directions". Business & Information Systems Engineering, May 2009.

[4] A. Gill, D. Banker, and P. Seltsika, "Moving Forward: Emerging Themes in Financial Services Technologies Adoption", Communications of the Association for Information Systems: Vol. 36, Article 12, 2015.

[5] Q. Zhang, L. Cheng, and R. Bautaba, "Cloud computing: State-of-the-art and Research challenges", J Internet Serv Appl 1:7-18, 2010.

[6] F. Paraiso, N. Haderer, P. Merle, R. Rouvoy, and L. Seinturier, "A Federated Multi-Cloud PaaS Infrastructure", IEEE Fifth International Conference on Cloud Computing, 2012.

[7] A. Velte, R. Elsenpeter, and T. J. Velte, "Cloud Computing: A practical approach". Tata McGraw-Hill Education Pvt. Ltd, 2009.

[8] U. Z. Rehman, F. K. Hussain, and O. K. Hussain, "Towards Multi-Criteria Cloud Service Selection", Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011.

[9] J. Bohli, N. Gruschka, M. Jensen, L.L. Iacono, and N. Marnau, "Security and Privacy-Enhancing Multi cloud Architectures", IEEE Transactions on Dependable and Secure Computing, Vol. 10, No' 4, 2013.

[10] E. Weintraub and Y. Cohen, "Cost Optimization of Cloud Computing Services in a Networked Environment", (IJACSA) International Journal of Advanced Computer Science and Applications ,Vol. 6, No. 4, pp. 148-157, 2015.

[11] E. Weintraub and Y. Cohen, "Optimizing User's Utility from Cloud Computing Services in a Networked Environment", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 10, pp. 153-163, 2015.

[12] A. Gill, D. Banker, P. Seltsika, "Moving Forward: Emerging Themes in Financial Services Technologies Adoption", Communications of the Association for Information Systems: Vol. 36, Article 12, 2015.

[13] Z. Chen, F. Han, J. Cao, X. Jiang, S. Chen, 'Cloud Computing-Based Forensic Analysis for Collaborative Network Security Management System', Tsinghua science and technology, Vol 18/1, 2/ 2013.

[14] L. Yung-Ming, C. Chia-Ling, "Analyzing The Pricing Models For Outsourcing Computing Services". PACIS Proceedings, 2012.

[15] H. Yang, M. Tale, "A Descriptive Literature Review and Classification of Cloud Computing Research". Communications of the Association for Information Systems: Vol. 31, Article 2, 2012.

[16] M. Walterbusch, B. Martens, F. Teuteberg, "Evaluating cloud computing services from a total cost of ownership perspective". Management Research Review Vol. 36 No. 6, pp. 613-638, 2013.

[17] S. El Kihal, C. Schlereth, B. Skiera, "Price comparison for Infrastructure-as-a-Service". In: ECIS Proceedings, 2012.

[18] S. J. Ward, A. Barker, "Observing the clouds: a survey and taxonomy of cloud monitoring". Journal of Cloud Computing, **3**:24, 2014.

[19] B. Blau, D. Neumann, C. Weinhardt, W. Michalk, "Provisioning of service mashup topologies", In: Proceedings of the 16th European conference on information systems, Galway, 2008.

[20] U. Z. Rehman, F. K. Hussain, O. K. Hussain, "Towards Multi-Criteria Cloud Service Selection", Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011.

[21] M. Al-Roomi, S. Al-Ebrahim, S. Buqrais, I. Ahmad, "Cloud Computing Pricing Models: A Survey". International Journal of Grid and Distributed Computing: Vol 6. No 5, 2013.

[22] G. Bitran, R. Caldentey, "An overview of pricing models for revenue management", Manufacturing & Service Operations Management 5(3):203–229, 2003.

[23] K. Lai, "Markets are dead long live markets", In: SIGecom Exchanges 5(4): pp 1–10, 2005.

[24] C. Weinhardt, B. Blau, J. Stößer, "Cloud Computing – A Classification, Business Models, and Research Directions". Business & Information Systems Engineering 05/2009.

[25] T. Pueschel, A. Anandasivam, S. Buschek, D. Neumann, "Making money with clouds: Revenue optimization through automated policy decisions". ECIS - European Conference on Information Systems 17, 2009.

[26] M. Lilienthal, "A Decision Support Model for Cloud Bursting", Business & Information Systems Engineering 2013.

[27] M. Walterbusch, B. Martens, F. Teuteberg, "Evaluating cloud computing services from a total cost of ownership perspective", Management Research Review Vol. 36 No. 6, pp. 613-638, 2013.

[28] P. Koehler, A. Anandasivam, M. Dan, C. Weinhardt, "Customer heterogeneity and tariff biases in cloud computing", Thirty First International Conference on Information Systems, St. Louis, ICIS proceedings, 2010.

[29] J. K. MacKie-Mason, H. R. Varian, "Pricing Congestible Network Resources". IEEE Journal on Selected Areas in Communications 13, number 7, 1995.

[30] H. R. Varian, "Economics of Information Technology". Working Paper, 2003.

[31] P. Koehler, A. Anandasivam, and A. Dan, “Cloud services from a consumer perspective”, AMCIS 2010 Proceedings, 2010.

[32] P. E. Green, A. M. Krieger, and Y. J. Wind, “Thirty Years of Conjoint Analysis: Reflections and Prospects”. Interfaces Vol. 31, No.3, 2001.

[33] M. A. Armbrust, R.Fox, A.J. Griffith, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia., "Above the Clouds: A Berkeley View of Cloud Computing”. Technical Report, Berkeley: CA, 2009.

[34] Z. Chen, F. Han, J. Cao, X. Jiang, and S. Chen, “Cloud Computing-Based Forensic Analysis for Collaborative Network Security Management System”. Tsinghua science and technology, Vol 18, No. (1, 2), 2013.

[35] W .Venters, and E.A.Whitley, “A critical review of cloud computing: researching desires and realities”, Journal of Information Technology, Vol. 27, No.3, 2012.

[36] P. Koehler, A. Anandasivam, M. Dan, and C. Weinhardt, “Customer heterogeneity and tariff biases in cloud computing”. Thirty First International Conference on Information Systems, St. Louis, ICIS proceedings, 2010.

[37] R. Weiber, and D. Mühlhaus, “Auswahl von Eigenschaften und Ausprägungen bei der Conjointanalyse”. In “Conjointanalyse”, by D. Baier and M. Brusch, Heidelberg: Springer, 2009.

[38] K. Bogataj, A. Pucihar, “Business Model Factors Influencing Cloud Computing Adoption: Differences in Opinion”, BLED Conference Proceedings, Bled:Slovenia, 2013.

[39] E. Weintraub, Y. Cohen, "Security Risk Assessment of Cloud Computing Services in a Networked Environment", International Journal of Advanced Computer Science and Applications (IJACSA), 7, (11), 2016.

[40] R. Latif, H. Abbas, S. Assar, & Q. Ali, "Cloud computing risk assessment: a systematic literature review". In Future Information Technology, Springer Berlin Heidelberg, 2014.

[41] E. Furuncu, & I. Sogukpinar, "Scalable risk assessment method for cloud computing using game theory", (CCRAM). Computer Standards & Interfaces, 38, 2015.

[42] T. Ackermann, T. Widjaja, A. Benlian, and P. Buzmann, "Percieved IT Security Risks of Cloud Computing: Conceptualization and Scale Development", Thirty Third International Conference on Information Systems, Orlando USA, 2012.

[43] S. B. Yadav, and D. Tianxi, "A Comprehensive Method to Assess Work System Security Risk," Communications of the Association for Information Systems: Vol. 34, Article 8, 2014.

[44] D. A. B. Fernandes, L. F. B. Soares, J. V. Gomes, M. M Freire and P. R. M. Inácio, "Security issues in cloud environments: a survey", Int. J. Inf. Secur. 13:113–170, 2014.

[45] B. Mansukhani and T. A. Zia, "The Security Challemges and Countermeasures of Virtual Cloud", Australian Information Security Management Confference, 2012.

[46] ISACA, "Security Considerations for Cloud Computing", USA, 2012.

Ant Colony Optimization (ACO) based Routing Protocols for Wireless Sensor Networks (WSN): A Survey

Anand Nayyar
Ph. D (Computer Science) Research Scholar
Desh Bhagat University, Mandi Gobindgarh

Rajeshwar Singh
Director, Doaba Group of Colleges, Nawanshahr

***Abstract*—Wireless Sensor Networks have several issues and challenges with regard to Energy Efficiency, Limited Computational capability, Routing Overhead, Packet Delivery and many more. Designing Energy Efficient Routing Protocol has always been a limiting factor cum issue for Wireless Sensor Networks. Varied routing protocols being proposed till date for overcoming these issues based on Swarm Intelligence. Swarm Intelligence is concerned with study of combined behavior of systems designed by varied components for coordinating among themselves via decentralized controls and self-organization. Algorithms based on Swarm Intelligence, nature based intelligence are highly robust, adaptive and scalable. This paper presents comprehensive survey of Ant Colony Optimization based routing protocols for Wireless Sensor Networks to give better platform for researchers to work on various shortcomings of protocols developed till date to develop efficient routing protocol for WSN in near future.**

Keywords*—Wireless Sensor Networks; Routing; Routing Protocols; Swarm Intelligence (SI); Ant Based Routing; Ant Colony Optimization

I. Introduction

In 21st Century, Wireless Sensor Networks (WSN) is regarded as one of the fastest evolving technologies. Wireless Sensor Network is enabled by recent sophisticated advancements in MEMS and Wireless communication technologies. A Wireless Sensor Network infrastructure is composed of nodes working autonomously with sensing, computation and networking elements which enables end user to measure, monitor and act to various events or phenomenon in different environments [1,2]. Wireless Sensor Networks features tiny, compact-size, and inexpensive computational wireless transmitting nodes with limited energy scattered over the area with diverse parameters like store and forward the data to central location for further processing. WSN nodes have the ability to operate in any sort of environment and network with other nodes to carry out data transmission tasks. WSNs can be easily deployed in various applications including civil, military, environmental monitoring, surveillance, health care applications, industrial production, transportation, space technology and more.

The implementation of Wireless Sensor nodes is done in ad-hoc manner without any sort of appropriate planning or research. Once deployed in real-world, the sensor network with other nodes autonomously transmit data back to sink node using different routing protocols. WSN nodes are battery powered, so limited energy is always a barrier, as it becomes difficult to change or recharge the batteries of sensor nodes in live operational environments [3].

Sensor nodes have limited transmission range and both data and control packets have to route from node to node back to sink node only via multi-hop fashion way [39]. Despite of varied objectives of sensor network applications, the primary function of WSN node is to collect specific data from the environment, process and transmit the data back to sink node via radio transmitter. In order to facilitate effective transmission of nodes among each other, the routing protocols should be highly efficient. A large number of challenges and issues in terms of theory and practical are taken into account to make the routing protocols efficient like maximization of network life-time, efficient route-discovery as nodes are autonomous which means that protocols must be self-organizing and protocols must handle random and complex environments and all sorts of radio interference to easily discover and maintain efficient Multihop routing paths [4].

Traditional or Classical routing protocols of Wireless Sensor Networks (WSNs) are not energy efficient and scalable to face varied challenges in dense and complex environments like military, production, medical and many other scenarios. So, there is an emergent need of new routing protocol based on Swarm Intelligence based technique for performing efficient routing and maintaining energy efficiency among nodes during data transmission. Till today, enormous types of routing protocols have been proposed for WSN based on varied social insects like Ants, Birds, Cats, Dog, Bats, Elephants on basis of their real world working mechanisms and optimizations. The goal of this research paper is to review the most prominent routing protocols based on Ant Colony Optimization (ACO) for WSN.

Swarm Intelligence [5-9] is a novel area considered for development of various optimizations in diverse areas. Currently, SI is regarded as important and foremost choice for every researcher across nook and corner in the world to develop efficient routing protocols for WSN/MANETS/VANETS/FANETS and many more areas.

Swarm Intelligence is a group of homogenous individual agents, having capability of self-organization, interaction among themselves and with environment.

Swarm Intelligence is originally defined as “Any Attempt to design algorithms or distributed problem-solving devices inspired by collective behavior of social insects and other animal societies”.

Swarm Intelligence was first conceptualized by G.Beni, Hackwood and J.Wang in 1989. Swarm Intelligence is primarily focused on study of integrated behavior of social insects as well as other animal societies using decentralized controls and self-organization. Swarm Intelligence enables tremendous increase in performance, robustness, scalability and efficiency by providing solution to complex problems. SI being an important concept in field of Artificial Intelligence and Computer Science focusses on development of various algorithms to deploy many simple agents with no rule and leading to global behavior using Ants, Birds, Honey bees etc. The interactions between the individual swarms can be direct or indirect. Direct interactions can be done via mode of audio or Video. Example: Birds interact among each other via making specific sounds and Bees interact with their communities using Waggle Dance. Indirect interactions mean interaction via environment i.e. One Swam Agent change the environment and other swarm responds to the changing Environment. Example: Ant Colony Optimization (ACO).

Swarm Intelligence based Techniques have widened scope in area of WSN. The most widely used techniques are Ant Colony Optimization and Bee Colony Optimization as the routing protocols based on ACO or Bee Colony are highly efficient in varied parameters like energy, robustness, scalability thus perform better in complex transmission environments.

A. Organization of Paper

The rest of the paper is organized as follows. Section 2 summarizes various Design challenges cum issues faced in designing routing protocols for WSNs. Section 3 briefly overviews the concept of Ant Colony Optimization-Metaheuristic, Algorithm and Implementation with Wireless Sensor Network. Section 4, presents detailed review of various selected routing protocols based on Ant Colony Optimization for Wireless Sensor Networks. Finally, Section 5 concludes the paper with directions for future research.

II. Challenges During Design for Routing Protocols oF Wireless Sensor Networks

A tremendous research has been carried out by researchers to overcome various challenges in WSN for successful real world operations.

WSNs being composed of tiny nodes having limited memory, low-processor, less-energy and small-bandwidth capabilities in-turn face strict constraints while designing routing protocols. WSNs transmit voluminous amounts of data back to sink node which makes use of energy, bandwidth and computing power. As sensor nodes operate in dynamic environment, large number of challenges are also faced in terms of architecture.

So, the utmost requirement is to design energy efficient routing protocols to overcome various additional challenges like: Power, Mobility, Connectivity, QoS, Data Aggregation, Deployment, Security, Cost, Congestion, Latency, Localization and many more.

The following are some of the issues which are to be taken into serious consideration by researchers to design Efficient Routing Protocol for WSN [2, 10, 11, 12, 13]:

- Less Computation and Limited Memory Requirements: Sensor nodes comprising Wireless Sensor Networks are equipped with Low-power CPU and tiny Memory in KBs like ATMega Processors containing 16KB to 128KB memory. So, Routing protocols should be designed consuming less CPU-power and execution should be feasible in limited amounts of memory.

- Deployment: Sensor nodes being deployed in random manner without any planning or research and are application dependent which ultimately affects the overall performance of routing protocol. So, in order to deal with ad-hoc environment, the routing protocol should be competent enough to self-organize the nodes and establish the paths among each other for data transmission and energy consumption.

- Energy Efficiency: The primary challenge for every routing protocol deployed in sensor nodes is Energy Efficiency. Sensor nodes exhaust the energy while performing varied tasks like Sensing, Computation and Transmission. It becomes utmost important that routing protocol should (1): Discover effective paths among nodes for transmission among each other and sending back the data to sink node. (2): Allocate the forwarding of data packets across multiple paths i.e. Effective multi-path routing should be integrated. (3): Activation of only those nodes used in transmission and keeping non-transmitting nodes in sleep mode.

- Scalability: As sensor network comprise of hundreds to thousands of nodes deployed for sensing the environment, mark utmost requirement i.e. routing protocol should be scalable enough to handle and respond to diversified events.

- Fault Tolerance: Lots of issues can occur in real world to sensor nodes in terms of power failure, physical damage, radio interference which overall affects the entire WSN network performance. Routing protocols should be fully compliant and compatible to handle these issues and should generate new routes among nodes to sink nodes in case of any node failure in between so that transmission can work without any sort of hic-cups.

- Latency/End-To-End Delay: Latency, in simple terms, is regarded as the time taken by a data packet to reach from node to sink node. Latency is measured in: One-Way: The time from source to sink, Round-Trip: The one-way latency from source to sink and from sink back to the source. Apart from this, latency can also be caused by Multi-Hop relays and data aggregation. So, routing protocol should be fully efficient and should take less time in transmission of data from nodes to sink nodes which also reduces the problem of congestion and packet failure in overall network operation.

- Quality of Service (QoS): Another challenge in effective design of routing protocol is Quality of Service (QoS). While designing routing protocol, QoS parameters like Jitter, Bandwidth, Delay and Reliability should be considered so that various aspects like data reliability, energy efficiency, collaborative processing can be maintained in WSN network.
- Data Gathering and Aggregation: Data aggregation and gathering can be event-driven, query-driven, continuous or hybrid combination. Data gathering methods occupy significant position in WSN routing, as after getting data, the node has to transmit the data back to sink node.

In lieu of various issues discussed above, researchers have proposed different protocols for WSN with regard to routing optimization as routing in WSN, till date, is surrounded by lots of challenges and constraints. Proposed routing protocols consider various sensor node characteristics with regard to application and architecture.

III. Nature based Routing-Ant Colony Optimization-Overview and Algorithm

Mostly the routing protocols being proposed and developed by researchers for Wireless Sensor Networks based on SI are based on Ants & Bee Colonies. The foraging behavior of insect societies is regarded as the major source of backbone to design highly energy efficient and sophisticated routing protocols for WSN.

Ants, during process of foraging, collectively explore the environment to find traces of food sources, and once food source is located, ant's setup paths between the nest and food sources to effectively transport the food back to nest. Therefore, the collective foraging behavior of ants includes environmental exploration, discovery of route, setting and use of highly efficient routing paths.

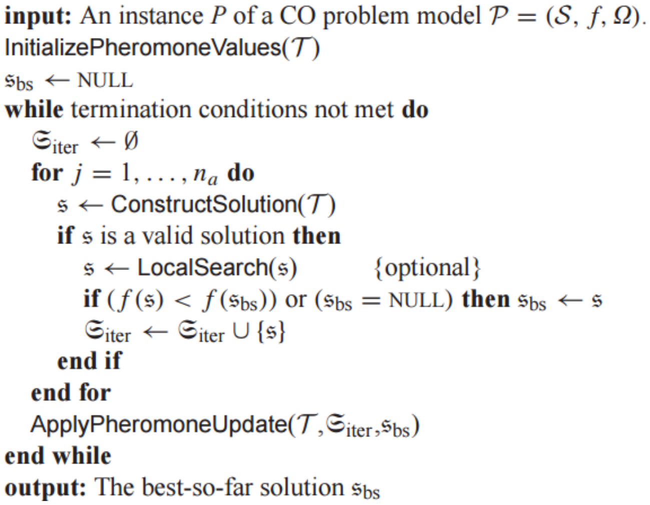
Ant Colony Optimization [14] (ACO) is the most utilized optimization technique in swarm intelligence for approximation determination. ACO belongs to class of metaheuristics which are regarded as Approximation Algorithms and lay foundations for obtaining highly efficient solutions to CO problems in timely manner.

ACO, algorithms considering the optimization and efficiency has become an important source of foundation for researchers to develop algorithms for routing protocols for Wireless Sensor Networks. ACO, apart from WSN has also been applied to various other engineering disciplines and other areas of computer science for solving complex problems and determining optimized solutions.

A. Ant Colony Optimization- Computational SI Technique-General Working and Algorithm

Ant Colony Optimization (ACO),[14-22] was discovered by M. Dorigo and colleagues for finding solutions to varied Hard CO problems in early 1990s. The basic foundation of ACO algorithms are real ant colonies. Ants roam randomly in the environment to determine food source and find the shortest path between food source and nest. In order to exchange information regarding which path to follow, ants communicate via use of chemical substance called Pheromone. As ants move from nest to food source, lay a trail of pheromone and other ants follow the same trail, laying trail of pheromone. The trail becomes more attractive when followed by huge majority of ants. Using this mechanism, ants are able to transport the food from source to nest in an efficient way.

Ant Colony Optimization Algorithm



```
input: An instance $P$ of a CO problem model $\mathcal{P} = (\mathcal{S}, f, \Omega)$.
InitializePheromoneValues($\mathcal{T}$)
$\mathfrak{s}_{bs} \leftarrow$ NULL
while termination conditions not met do
    $\mathfrak{S}_{iter} \leftarrow \emptyset$
    for $j = 1, \ldots, n_a$ do
        $\mathfrak{s} \leftarrow$ ConstructSolution($\mathcal{T}$)
        if $\mathfrak{s}$ is a valid solution then
            $\mathfrak{s} \leftarrow$ LocalSearch($\mathfrak{s}$)        {optional}
            if $(f(\mathfrak{s}) < f(\mathfrak{s}_{bs}))$ or $(\mathfrak{s}_{bs} =$ NULL$)$ then $\mathfrak{s}_{bs} \leftarrow \mathfrak{s}$
            $\mathfrak{S}_{iter} \leftarrow \mathfrak{S}_{iter} \cup \{\mathfrak{s}\}$
        end if
    end for
    ApplyPheromoneUpdate($\mathcal{T}$,$\mathfrak{S}_{iter}$,$\mathfrak{s}_{bs}$)
end while
output: The best-so-far solution $\mathfrak{s}_{bs}$
```

B. ACO and Wireless Sensor Networks

Wireless Sensor Networks (WSN), comprising thousands of autonomous and limited energy sensor nodes are deployed in wide range of environments. Efficient Routing is still demanding area to be researched out for developing efficient routing protocols for WSN.

Ant Colony Optimization based Algorithms comprise of varied distinguished features as listed below which makes it the most suitable for developing routing algorithms for Wireless Sensor Networks [23]:

1) As ACO algorithms are fully distributed, so failure rates are reduced to large extent in sensor nodes communications.

2) Simple operations can be performed in each and every node for routing of packets among nodes and back to sink node.

3) Autonomous integration of ants, and the algorithms are based on Agent's Synchronous.

4) ACO algorithms have capability of Self-Organizing which is very important as sensor nodes, when deployed randomly have to fully robust, scalable and fault tolerant. ACO algorithms makes WSN networks fully self-organization compliant.

5) ACO algorithms are very well suited to adapt to all kinds of changes in real-world topology and increase in number of nodes and packet traffic.

6) ACO algorithms solve complex CO problems, making them well suitable for highly complex situations when sensor nodes are deployed especially in Real-Time Monitoring, Production and Military based Battlefield's monitoring.

IV. Review of Routing Protocols for WSN based on Ant Colony Optimization [23-30, 38, 40]

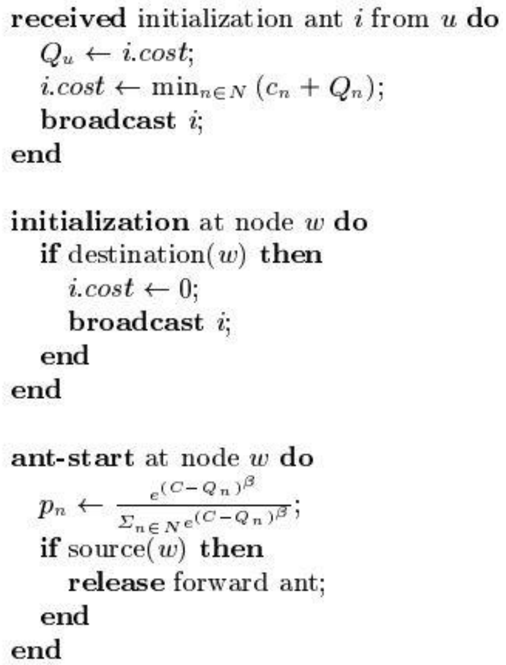
In this section, Ant Colony Optimization based Routing Protocols for WSNs are highlighted.

A. Sensor Driven Cost-Aware Ant Routing(SC) [31]

The main problem surrounding all Basic Ant Routing Algorithms is that all the forwarding ants normally consume lots of time to locate the destination, even when a tabu list is being utilized (i.e. Repeating nodes are not included). This situation usually occurs when ants primarily don't have any idea regarding the exact destination. Only when the destination is located, the links are traversed along with certain probabilities of link exchange.

In SC Routing, the routing performance is improvised; it is assumed that forward ants are equipped with sensors to locate the best destination for food at the initial process of routing. In addition to smart sensing ability of ants, each node stores the probability distribution and every node estimates and stores the cost to the destination from neighboring nodes. It suffers from redundant data when obstacle arises in path leading to sensing errors.

SC Algorithm:



```
received initialization ant $i$ from $u$ do
    $Q_u \leftarrow i.cost$;
    $i.cost \leftarrow \min_{n \in N} (c_n + Q_n)$;
    broadcast $i$;
end

initialization at node $w$ do
    if destination($w$) then
        $i.cost \leftarrow 0$;
        broadcast $i$;
    end
end

ant-start at node $w$ do
    $p_n \leftarrow \frac{e^{(C-Q_n)^\beta}}{\Sigma_{n \in N} e^{(C-Q_n)^\beta}}$;
    if source($w$) then
        release forward ant;
    end
end
```

B. Energy Efficient Ant Based Routing (EEABR) [32]

Energy Efficient Ant Based Routing (EEABR) algorithm, proposed by T. Camilo et al [32] is an improvised routing protocol based on Ant Colony Optimization (ACO) metaheuristic. The protocol was designed with an objective to enhance sensor nodes energy by reducing communication overhead in discovering the paths from source to destination. The protocol adds new functionalities in pheromone tables updation of sensor nodes.

Algorithm

1) In EEABR routing protocol, at regular interval period of time, from each network node, a forward ant is launched to determine a path from nest to food source. The identifier of every visited node is saved in memory and carried forward by ant. Each network node has routing table with N entries, one for each possible solution, and destination is one of the entry in nodes routing table.

2) At every node, the ant selects the next hop using the same ACO metaheuristic probabilistic rule.

3) When the forward ant reaches the food destination, it is transmitted back to proceeding ant, whose main task is to update the pheromone trail of the path used by forward ant to reach from nest to source and also stored in memory.

4) The destination node computes the amount of pheromone trail that the ant will drop during the journey, before backward ant starts the journey.

5) When the node, receives the backward ant coming from neighboring node, it updates the routing table.

6) When the backward ant reaches the nest, the actual path is determined by other ants to follow.

Simulation of EEABR with BABR (Basic Ant Based routing algorithm) and IABR (Improvised Ant-Based Routing Algorithm) is done on NS-2 simulator on varied parameters like Average Energy, Minimum Energy, Standard Deviation and Energy Efficiency and overall EEABR performs much better as compared to other two routing protocols. The only drawback of EEABR is lack of QoS and somewhat delay in packet delivery.

C. Flooded Forward Ant Routing (FF) [31]

Flooded Forward Ant Routing (FF) was developed to overcome the shortcomings of misguiding paths due to obstacles in SC protocol even when ants are equipped with sensors. When the exact destination is unknown at the beginning by ant and even the cost cannot be determined, SC protocol was reduced to Basic Ant Routing and still the problem of unknown wandering around the network by ant to find the destination exist. In that case, FF protocol was introduced to remove the problem.

FF protocol exploits the network via broadcast channel of WSN which means FF protocol makes use of Broadcast method of sensor networks to route the network packets from source to destination. The objective is to flood forward ants to the destination. If the food search is successful, forward ants will direct backward ants to traverse backwards to the source. Multiple paths are updated by one flooding phase and probabilities are updated in the same manner as in Basic Ant Routing Protocol.

D. Flooded Piggyback Ant Routing (FP) [31]

In flooded Piggyback Ant Routing (FP), a novel specimen of ants i.e. Data Ants was introduced. The forward list is carried by FP. In FP protocol, forward ants and data ants are combined via constrained flooding to route data packets and search for energy efficient paths in the network.

FP protocol was compared with SC, FF and Basic ACO routing protocols in RMASE (Routing Modeling Application Simulation Environment) simulator. Results showed FP is not an energy efficient routing protocol. FF protocol is efficient in reducing delay and SC remains highly energy efficient routing protocol among FF, FP and Basic ACO routing protocol.

E. Energy-Delay Ant Based (E-D Ants) [33]

Energy-Delay Ant Based (E-D Ants) was proposed by Wen et. al (2008). E&D ants is a reactive routing protocol being based on ant algorithms for performing varied routing operations. E-D Ants Protocol is based on Energy*Delay metrics to enhance network lifetime and minimize propagation delay by making use of a novel variation of Reinforcement Learning (RL).

The Mathematical expression of E-D Ants Protocol is:

$$g(t) = \min(\text{Energy} * \text{Delay}) \qquad (1)$$

The protocol works on Iterative generation and unicast transmission of multiple forward ants to minimize energy and delay like AntNet Protocol. In this protocol, every ant stores the residual energy level and hop delay experience in its stack moving from node to node.

E-D Ants Routing protocol was simulated in OPNET Simulator using 50 sensor nodes in area of 100x100 m and compared with two routing protocols: AntNet and AntChain on basis of Energy Efficiency, Delay and Routing Overhead. The results showed E-D Ants Protocol is almost 150% efficient as compared to other two protocols. E-D Ants protocol is also efficient routing protocol in determining optimal paths from source to destination.

F. Ant Colony Based Reinforcement Learning Algorithm (AR and IAR) [34-35]

Adaptive Routing (AR) and Improved Adaptive Routing (IAR), proposed by Ghasemaghaei et. al (2007) uses probability distribution like other Ant-Colony based routing protocols in finding optimal paths from source to destination. The only difference between AR and IAR with other ACO based routing protocols is the use of reinforcement learning algorithm by backward ants to get efficient routing path from source to destination.

In AR and IAR, two types of ants are deployed:

1) Forward ant (F_{ant})- travelling from source node (s) to destination node (d)

2) Backward ant (B_{ant}): which is generated by F_{ant} when F_{ant} reaches the destination d.

The backward ant gets back to sink node via information supplied by forward ant. But backward ant makes use of reinforcement learning method to get better and most optimal route as compared to the route being chosen by forward ant and updates the routing table of sensor nodes visited during reverse journey.

AR and IAR algorithms were simulated on Java based simulator using 7x7 sensor node grid for 200 seconds. AR and IAR algorithms are compared with 4 Routing Algorithms: Basic Ant Routing, SC Ant Routing, FF and FP Routing Algorithm on parameters like Latency, Energy Consumption, Success Rates. Simulation results showed AR and IAR much efficient in every parameter as compared to other 4 routing protocols.

G. Basic Ant Based Routing(BABR) for Wireless Sensor Networks (WSN) [17] [21]

Ant Colony Optimization (ACO), is nature-inspired metaheuristic for solving complex Combinatorial Problems (CO). The main component of ACO algorithm is Pheromone Model.

ACO, being an optimization approach, used to solve complex problems by iterating the following two steps:

1) Using a Pheromone model, that is, a parametrized probability distribution over the solution space;

2) The candidate solutions are used to modify the pheromone values in a way that is deemed to bias future sampling forward high quality solutions.

Basic Ant Based Routing leads to the development of AntNet Algorithm which can be summarized as follows:

a) Forward ant is launched from source node to sink node to determine the optimal path to destination.

b) The main task of forward ant is to locate the food source with equal probability by using neighboring nodes with minimum cost joining its source to sink.

c) As ants move forward from node to node to reach the destination, the routing table gets updated side by side.

d) Forward ants calculate all the information about the time length, congestion status and the node identifiers of the followed path.

e) On reaching the destination node, the backward ant is created which follows the same path as forward ant, in opposite direction i.e. from food source to nest.

f) During backward travel, local models of the network status and the local routing table of each visited node are modified by the agents as a function of the path they followed and of its goodness.

H. Ant Based Quality of Service Routing (ACO-QoSR) [36]

ACO-QoSR, a reactive routing algorithm was developed by Cai et. al in 2006 to tackle problems of constraint delay and energy in Wireless Sensor Networks. The basic objective behind the development of ACO-QoSR routing is to find optimal routes between varied sensor nodes to sink node in such a way that the total end-to-end delay is less than a boundary value, while the energy residual ratio i.e. $ERR=E_{residual}/E_{initial}$ is above a certain value.

ACO-QoSR Algorithm

When source node wants to send data, it first checks routing table to determine optimal path. Route probing will only start if there are no unexpired paths to the destination, and node needs to cache data waiting for transmission at the same time. Forwards ants does the task for route probing and after route discovery cached data is sent to destination in no time. In order to reduce time delay of route discovery, ACO-QoSR algorithm starts a full route probe phase at the time of network initialization.

Forward Ant Phase: In forward ant's phase, if the sending/source sensor unable to find a favorable path to sink node in routing table, it will generate a number of forward ants to search for optimal paths to destination. Forward ants will establish pheromone track between source to destination node. Forward ants comprise of various parameters: Timestamp origin, source and destination address. The main aim of forward ant is to collect intermediate node's local information and record the path information of various nodes from source to destination.

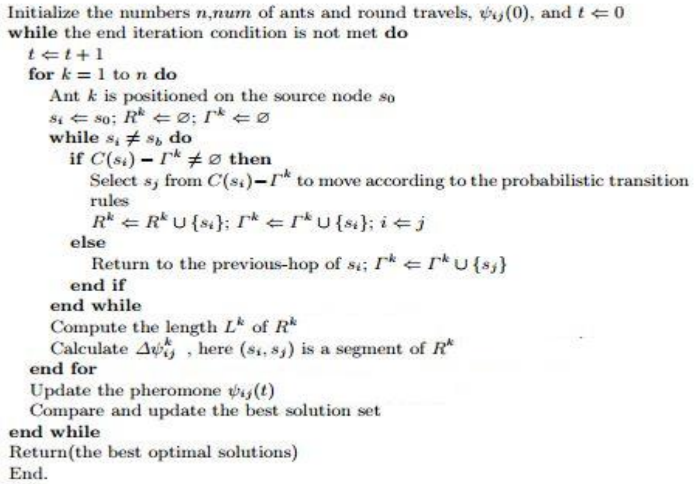
Backward Ants Phase: When the forward ant reaches the destination, the forward ant will be killed and backward ant will be generated which carries source and destination address, backward ant ID, path information from forward ant and pheromone update value.

Route Maintenance Phase: The entries in the routing table are basically pheromone values and probabilities that next-hop is a specific neighbor. Probabilities allow the ants to roam randomly in the environment and find new and optimal paths. Once the new optimal paths are discovered, the next hop probabilities are updated to routing table to reflect new paths from source nodes to sink nodes.

ACO-QoSR protocol was simulated in NS-2 Simulator [38] considering the network of 100 sensor nodes in 1000x1000m area. ACO-QoSR protocol was compared with AODV and DSDV protocols on parameters like end-to-end delay, packet delivery ratio, routing overhead and path's normalized energy residual ratio. Simulation results showed that ACO-QoSR has better energy residual ratio and less overhead but packet delivery ratio is just average as compared to AODV and DSDV and routing overhead is small.

I. Ant Colony Optimization based Location-aware Routing (ACLR) [37]

Ant Colony Optimization based Location Aware Routing (ACLR), a High Performance Routing Protocol for Wireless Sensor Networks was designed by Wang et. al in 2008. The principle behind the working of ACLR protocol is determination and selection of next hop by ants to a subset of the set of the neighbors of the current node which guarantee for the packet delivery rather than searching of whole neighbors to avoid loops. The protocol also determines the amount of pheromone which laid by the ant from source node to sink node. In addition, the protocol also proposes a novel scheme to evaporate the pheromone on the different segments of a certain route as per residual energy and the location information of nodes.



```
Initialize the numbers $n$,$num$ of ants and round travels, $\psi_{ij}(0)$, and $t \Leftarrow 0$
while the end iteration condition is not met do
  $t \Leftarrow t+1$
  for $k = 1$ to $n$ do
    Ant $k$ is positioned on the source node $s_0$
    $s_i \Leftarrow s_0$; $R^k \Leftarrow \varnothing$; $\Gamma^k \Leftarrow \varnothing$
    while $s_i \neq s_b$ do
      if $C(s_i) - \Gamma^k \neq \varnothing$ then
        Select $s_j$ from $C(s_i) - \Gamma^k$ to move according to the probabilistic transition
        rules
        $R^k \Leftarrow R^k \cup \{s_i\}$; $\Gamma^k \Leftarrow \Gamma^k \cup \{s_i\}$; $i \Leftarrow j$
      else
        Return to the previous-hop of $s_i$; $\Gamma^k \Leftarrow \Gamma^k \cup \{s_j\}$
      end if
    end while
    Compute the length $L^k$ of $R^k$
    Calculate $\Delta\psi_{ij}^k$ , here $(s_i, s_j)$ is a segment of $R^k$
  end for
  Update the pheromone $\psi_{ij}(t)$
  Compare and update the best solution set
end while
Return(the best optimal solutions)
End.
```

ACLR Algorithm

ACLR Algorithm was simulated on OPNET Simulator and compared with 4 algorithms: Basic Ant Routing (BAR), SC, FP and IAR using network area of 200x300 m and 10000 sensors. Performance of ACLR algorithm is determined on energy consumption, efficiency and packet delivery latency. Results showed that ACLR consumes less energy as compared to 4 other algorithms and it is also better in terms of Packet Delivery.

V. Conclusion and Future Directions

Wireless Sensor Networks, being strongest platform for research across wide forum of researchers around the world. Wireless Sensor Network nodes are resource-constrained nodes and lots have to be done regarding improvement of various parameters to make WSN network more adaptable in real world.

The design and development of energy efficient, robust, scalable and effective packet delivery routing protocol in WSN network is a challenging task. Diverse optimization fields like Swarm Intelligence, Fuzzy Logics, Genetic Algorithms are being utilized by researchers to develop routing protocols. One of the most utilized novel domain in development in WSN routing protocols is Swarm Intelligence. So, taking SI into consideration wide range of protocols are developed and still lots are under rapid development and testing phase by researchers, most specifically, taking two main techniques into consideration: Ant Colonies: Based on foraging behavior of Ants and Bee Colonies: Considering efficient way of communication of Bees.

In this paper, we have presented a detailed comprehensive review of Ant Colony Optimization based Routing Protocols for Wireless Sensor Networks.

Considering all the protocols in this paper, it is being observed that currently very little research is being done regarding Security, QoS parameters and most of the protocols assume that sink node is stationary which is a limited barrier towards research. Research should be done seriously considering each and every node in topology to be dynamic and mobile and having random changing scenario. New routing Protocols which are required in WSN should be able to handle mobility overhead and random topology changes by maintaining optimal paths in route discovery, selection and maintenance and especially strong consideration should be done in maintaining energy efficiency of each and every node in the network.

On other hand, apart from developing and testing the new routing protocols over Simulators and Testbeds, it is also recommended that research should be conducted on live sensor nodes for determining accurate performance on the basis of algorithm/protocol proposed.

We strongly believe that considering this paper, researchers would take SI into more serious consideration and come up with more advanced and efficient routing protocols well tested with diverse parameters and fully functional to be adaptable in real world sensor networks.

VI. Future Scope

In near future, considering the pros and cons of different routing protocols being developed for WSN using Ant Colony Optimization, a Novel Multipath based routing protocol more efficient in packet delivery, end-to-end delay, less routing overhead and Energy Efficiency will be developed.

References

[1] Akkaya, K., & Younis, M. (2005). A survey on routing protocols for wireless sensor networks. *Ad hoc networks*, *3*(3), 325-349.

[2] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer networks*, *38*(4), 393-422.

[3] Chong, C. Y., & Kumar, S. P. (2003). Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE*, *91*(8), 1247-1256.

[4] Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41-50.

[5] Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: from natural to artificial systems* (No. 1). Oxford university press.

[6] Kennedy, J., Kennedy, J. F., Eberhart, R. C., & Shi, Y. (2001). *Swarm intelligence*. Morgan Kaufmann.

[7] Engelbrecht, A. P. (2006). *Fundamentals of computational swarm intelligence*. John Wiley & Sons.

[8] Keerthi, S., Ashwini, K., & Vijaykumar, M. V. (2015). Survey Paper on Swarm Intelligence. International Journal of Computer Applications, 115(5).

[9] Di Caro, G. A. (2014). Principles of swarm intelligence for adaptive routing telecommunication networks. *Sistemi intelligenti*, *26*(3), 443-464.

[10] Raghavendra, C. S., Sivalingam, K. M., & Znati, T. (Eds.). (2006). *Wireless sensor networks*. Springer.

[11] Zheng, J., & Jamalipour, A. (2009). *Wireless sensor networks: a networking perspective*. John Wiley & Sons.

[12] Misra, S., Zhang, I., & Misra, S. C. (Eds.). (2009). *Guide to wireless sensor networks*. Springer Science & Business Media.

[13] Goyal, D., & Tripathy, M. R. (2012, January). Routing protocols in wireless sensor networks: a survey. In *2012 Second International Conference on Advanced Computing & Communication Technologies* (pp. 474-480). IEEE.

[14] Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. *IEEE computational intelligence magazine*, *1*(4), 28-39.

[15] Yaseen, S. G., & Al-Slamy, N. M. (2008). Ant colony optimization. *IJCSNS*, *8*(6), 351.

[16] Stützle, T. (2009, April). Ant colony optimization. In *International Conference on Evolutionary Multi-Criterion Optimization* (pp. 2-2). Springer Berlin Heidelberg.

[17] Dorigo, M., & Blum, C. (2005). Ant colony optimization theory: A survey. *Theoretical computer science*, *344*(2), 243-278.

[18] Dorigo, M., & Stützle, T. (2003). The ant colony optimization metaheuristic: Algorithms, applications, and advances. In *Handbook of metaheuristics* (pp. 250-285). Springer US.

[19] Blum, C. (2005). Ant colony optimization: Introduction and recent trends. *Physics of Life reviews*, *2*(4), 353-373.

[20] Sim, K. M., & Sun, W. H. (2003). Ant colony optimization for routing and load-balancing: survey and new directions. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *33*(5), 560-572.

[21] Nayyar, A., & Singh, R. (2016, October). Ant Colony Optimization—Computational swarm intelligence technique. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on* (pp. 1493-1499). IEEE.

[22] Maniezzo, V., & Carbonaro, A. (2002). Ant colony optimization: an overview. In *Essays and surveys in metaheuristics* (pp. 469-492). Springer US.

[23] Farooq, M., & Di Caro, G. A. (2008). Routing protocols for next-generation networks inspired by collective behaviors of insect societies: An overview. In *Swarm Intelligence* (pp. 101-160). Springer Berlin Heidelberg.

[24] Saleem, M., Di Caro, G. A., & Farooq, M. (2011). Swarm intelligence based routing protocol for wireless sensor networks: Survey and future directions. *Information Sciences*, *181*(20), 4597-4624.

[25] Zungeru, A. M., Ang, L. M., & Seng, K. P. (2012). Classical and swarm intelligence based routing protocols for wireless sensor networks: A survey and comparison. *Journal of Network and Computer Applications*, *35*(5), 1508-1536.

[26] Zengin, A., & Tuncel, S. (2010). A survey on swarm intelligence based routing protocols in wireless sensor networks. *International Journal of Physical Sciences*, *5*(14), 2118-2126.

[27] Ali, Z., & Shahzad, W. (2011, July). Critical analysis of swarm intelligence based routing protocols in adhoc and sensor wireless networks. In *Computer Networks and Information Technology (ICCNIT), 2011 International Conference on* (pp. 287-292). IEEE.

[28] Wang, C., & Lin, Q. (2008, June). Swarm intelligence optimization based routing algorithm for Wireless Sensor Networks. In *Neural Networks and Signal Processing, 2008 International Conference on* (pp. 136-141). IEEE.

[29] Geetha, R., & Srikanth, G. U. (2012). Ant Colony optimization based Routing in various Networking Domains–A Survey. *International Research Journal of Mobile and Wireless Communications*, *3*(01), 424-428.

[30] Nayyar, A., & Singh, R. (2016). A Comprehensive Review of Ant Colony Optimization (ACO) Based Energy-Efficient Routing Protocols for Wireless Sensor Networks.

[31] Zhang Y, Kuhn LD, Fromherz MPJ (2004). "Improvements on Ant Routing for SensorNetworks," Ant Colony, Optimization And Swarm Intelligence, Lecture Notes in Computer Science, 2004, 3172: 289-313.

[32] Camilo, T., Carreto, C., Silva, J. S., & Boavida, F. (2006, September). An energy-efficient ant-based routing algorithm for wireless sensor networks. In *International Workshop on Ant Colony Optimization and Swarm Intelligence* (pp. 49-59). Springer Berlin Heidelberg.

[33] Wen, Y. F., Chen, Y. Q., & Pan, M. (2008). Adaptive ant-based routing in wireless sensor networks using Energy* Delay metrics. *Journal of Zhejiang University SCIENCE A*, *9*(4), 531-538.

[34] GhasemAghaei, R., Rahman, A. M., Rahman, M. A., Gueaieb, W., & El Saddik, A. (2008, March). Ant colony-based many-to-one sensory data

routing in wireless sensor networks. In *2008 IEEE/ACS International Conference on Computer Systems and Applications* (pp. 1005-1010). IEEE.

[35] GhasemAghaei, R., Rahman, M. A., Gueaieb, W., & El Saddik, A. (2007, May). Ant colony-based reinforcement learning algorithm for routing in wireless sensor networks. In *2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007* (pp. 1-6). IEEE.

[36] Cai, W., Jin, X., Zhang, Y., Chen, K., & Wang, R. (2006, September). ACO based QoS routing algorithm for wireless sensor networks. In International Conference on Ubiquitous Intelligence and Computing (pp. 419-428). Springer Berlin Heidelberg.

[37] Wang, X., Li, Q., Xiong, N., & Pan, Y. (2008, October). Ant colony optimization-based location-aware routing for wireless sensor networks. In *International Conference on Wireless Algorithms, Systems, and Applications* (pp. 109-120). Springer Berlin Heidelberg.

[38] Nayyar, A., & Singh, R. (2015). A Comprehensive Review of Simulation Tools for Wireless Sensor Networks (WSNs). *Journal of Wireless Networking and Communications*, *5*(1), 19-47.

[39] Nayyar, A., & Sharma, S. (2014). A Survey on Coverage and Connectivity Issues Surrounding Wireless Sensor Network. *IJRCCT*, *3*(1), 111-118.

[40] Kumar, A., & Nayyar, A. Energy Efficient Routing Protocols for Wireless Sensor Networks (WSNs) based on Clustering.

An Investigation of Analytic Decision During Driving Test

Samir Ghouali
Fac of Technology, STIC Lab,
University of Tlemcen,
Algeria

Yassine Zakarya Ghouali
Fac of Economics, Business and
Management Sciences, POLDEVA
Lab, University of Tlemcen,
Algeria

Mohammed Feham
Fac of Technology, STIC Lab,
University of Tlemcen,
Algeria

***Abstract*—To examine the long-term causality between Cardiorespiratory Electromyography Galvanic signals for 17 drivers taken from Stress Recognition in Automobile Drivers database.**

Methods: Two statistical methods, co-integration to reveal an eventual existence of a long-term relationship between ECG (Electrocardiograph), EMG (electromyography), GSR (galvanic skin resistance), heart rate (HR) and respiration, well as the Application of the model of Granger causality.

Results: ECG shows certain dependence to EMG, GSR, heart rate and respiration. The results for ECG dependent suggest that an increase of 1% in EMG, FOOTGSR, HAND GSR, HR and RESPIRATION implies a variation of ECG which take a value respectively of 0.016248%, 0.007241%, 0.028366%, 0.000511% and 0.000110% in the within dimension based on the FMOLS (Fully Modified Ordinary Least Squares). With same way, the result for ECG suggest that an increase of 1% in EMG, FOOT GSR, HAND GSR, HR and RESPIRATION implies a variation of ECG which take a value respectively, of 0.065684%, 0.014534%, 0.032800%, 0.000304%, 0.005986% in the between dimension based on the same method. The results of panel Granger causality show a bi-directional relationship between ECG and FOOT GSR, HAND GSR and respiration signals, it must be noted as a unidirectional causality from EMG to ECG.

Conclusion: This study shows the long-term interaction between the bio signals, and reveal how the understanding of these interactions can help the doctors to understand the risks that may exist between these interactions. The main advantage of a multidimensional and multivariate model is to solve a multitude of problems that prevent doctors to treat the patients better and is not the case for studies in two dimensions.

***Keywords*—Panel Co-integration; Panel Granger Causality; FMOLS and DOLS Estimators; Cardiorespiratory electromyography galvanic signals**

I. Introduction

Actually, ECG signal is not independent of the other physiological signals, and medical explanations can attest to this. Thus, this concept of interactions between the physiological signals must be well formulated and analyzed. In this context, this article was devoted to the presentation of a strategy to analyze the interactions between biomedical signals in order to develop an approach to diagnosis.

The choice of the mathematical models, resulting from the physiological signals, for the characterization of such or such pathology becomes crucial. In order to widen the detection of these anomalies, we proposed a statistical analysis of the physiological signals to research factors of causality between them.

Traffic has become a major human activity; its development has resulted in huge infrastructure projects of communication used for a variety of uses, implementing very different vehicles. The industry associated with them has played and continues to play a significant economic role.

The evidence is that the main causes of accidents were related to a lack of driver vigilance. This lack of vigilance is, in fact, the result of many factors that are identified as inattention, drowsiness, and errors related to fatigue. Medication, drugs, alcohol, health accidents are all causes of accidents deserve to be treated specifically because they have become the main source of road accidents [1], (Algeria was ranked fourth in the Arab countries, in terms of road traffic accidents, with a heavy balance sheet that was estimated at 44,907 accidents resulted in the death of 4540 people and left 69,582 injured nationally [2]).

Many teams, for nearly twenty years, have been mobilized to better understand the origins of this drowsiness and to detect the occurrence as early as possible in order to take the necessary security provisions: braking, stopping.

In the automotive world, this goal is called "active safety" and complements the "passive safety" which aims to reduce the extent of damage in case of an accident.

The situation today is that many approaches have been explored, leading to important scientific and technological developments on-board sensors and diagnostic methods. However, we can say that these developments are not yet arrived at an operational stage, mainly because of diagnostic errors: false alarms, fault detection ... that remain present and hinder implementation confident developed devices. The analysis of this situation shows three emergencies:

- Continue the validation work required by multiplying the test
- Develop methods and tools for detecting
- Perfect the valuation procedures for accurate comparison of results with embedded diagnostics "reference" systems expertise from physiological signals.

It is on this last point that took our work on "Contributions of mathematical models to analyze short and long-term objective physiological signals, the vigilance of car drivers." It enriches existing analysis of biological signals in three ways:

- The accuracy of features extraction of the real signal and it is not our purpose in this work.
- Develop telemedicine applications to monitoring the health status of the patient
- process automation implemented in order to provide medical experts the possibility of multiple testing and diagnostics "referring" to monitor the couple "driver-vehicle," and anticipate the occurrence of a hazard sufficiently well characterized to trigger an alarm or automatically make a rescue maneuver [3].

The panel data models knew these twenty last years a very sharp enthusiasm. This passion resulted in a true explosion amongst academic work founded on the panel date models. The aspects of the transposition of time series problems to the panels are detailed in what follows.

Also named the structure with double dimensions, the study in panel brings information richer as that available in time series [35]. Indeed, it constitutes a particularly invaluable statistical source for the analysis of the dynamic behaviors.

The profit which results from all that, is the possibility of modelling more complex individual behaviors and dynamic alternatives. The distinction between the dynamic macro/micro effects Interactional, is done by the addition of a temporal dimension, It is one of the advantages to the recourse to the data of panel as Hsiao [36] indicates it. It is necessary to discuss two significant factors such as the decomposition of the total variability of the data and the increase amongst degrees of freedom which aims to decrease the collinearity between the data.

The panel date has a major asset concerning the number of data. To use a significant number of data increases the degree of freedom and decreases the collinearity between the variables.

The panel data thus provide the possibility of deducing the individual behavior while making use of the behaviors of the other individuals. The use of data of panel also makes it possible to reduce the frequent problems in time series of collinearities between the explanatory variables thanks to the possibility of introducing inter individual differences. These individual effects have the second advantage of being able to identify and take account of the unobservable effects.

Obviously, nothing is perfect and each model contains limits. Among the disadvantages of the panel of data, we can cite an incomplete panel, a panel says not rolled, problems of heteroscedasticity and/or autocorrelation of the random variations, but it is not always easy to correct or avoid these disadvantages.

For this configuration, we are leaning models on panel data, where we considered a double dimension: an individual dimension and a temporal dimension. Individual dimension represents the patient and temporal dimension represents the studied physiological signals.

The taking into account of biomedical data also generated an increase in the temporal dimension which results in a transposition of the questions usually asked into time series, such as the stationnarity, the non-linearity or the temporal stability of the relations.

In this context, the use of data panel models at the same time makes possible to combine the advantages of working on the panel data and solving the problems of nonlinearity, heterogeneity and temporal instability. More precisely, these models authorize the existence of dynamic individual distinct being able to evolve in time while taking account of asymmetries.

These changes are therefore an interesting solution to meet the new challenges posed by the use of panel data. However, this field is relatively recent what implies that certain current debates of time series as non-stationariness according to non-linearity is not posed yet in panel data.

The issue of using these changes for forecasting purposes is also addressed in this article.

This article initially proposed an outline of methods devoted to the principal tests of unit roots, of Co-integration on panel data, models of estimates as well as the use of causality within the meaning of Granger in panel. This research experienced a great development since work pioneers of Levin and Lin and is today the object of multiple applications at the empirical level. The theoretical framework, which is the base of any empirical study, brings contents of legitimacy to our problems, as it is used to clarify the concepts and makes it possible to define each concept.

We worked out the analysis of the physiological signals containing a mathematical model as a panel, pertaining to the same family of Granger. This model exploits Co-integration on data panel and allows the quantification short and long terms if it exists using estimators FM-OLS and DOLS. Within this framework, we brought an analytical study of the interactions between the physiological signals of the drivers of vehicles using a modeling of the interactions between the cardiorespiratory signals galvanic electromyographies.

The goal of these studies is to propose the direction of causality between the physiological signals as well as the quantification of the rates of causality if it exists in Panel. We carried out the analysis of each scenario according to the following stages:

- To check if there exists a long-term relation between the vital signals.
- To quantify the rate of convergence of this long-term relation.
- To define the direction of causality between the signals on the basis of causality of Granger as a Panel.
- To understand the impact of the signals on the long-term heart.

Numerous studies have been devoted to the evaluation of causality; several applications are omnipresent in areas ranging from the economy [4, 5], directed information theory in networks [6], brain imaging field [7], genetics [8] and especially the analysis of biological systems, with a very special emphasis on the neural field [9, 10], the study of cardiac signals [11, 12, 13, 33, 34] and cardiorespiratory interactions [14, 15].

In this article, we will look at studied panel data causality and panel Co-integration of a number of physiological signals, derived from the Stress Recognition in Automobile Drivers dataset [16] from the PhysioBank database [17], ,this approach was then applied to electrocardiogram (ECG), electromyography (EMG), galvanic skin resistance (GSR) measured on the hand and foot, heart rate (HR) and RESPIRATION.

This paper is organized as follows: In Section 2, we will establish the data used and the methodology. In section 3, we present the approach of Co-integration, in section 4, 5 and 6 the estimated long-term relationship, and Granger causality tests respectively. Finally, we lead an analysis, scientific discussion, conclusion and a projection of perspectives.

II. Variables, Data Analysis and Methodology

A. Variables

Electrocardiogram (ECG) is a recording of the electrical activity of the heart. The initial diagnosis of heart attack is usually made through observation of a combination of clinical symptoms and characteristic ECG changes [26].

Electromyography (EMG) is a diagnostic procedure to assess the health of muscles and the nerve cells that control them (motor neurons).

The Galvanic Skin Response (GSR) is defined as a change in the electrical properties of the skin. The measurement is relatively simple, and has a good repeatability. Therefore, the GSR measurement can be considered a simple and useful tool for examination of the autonomous nervous system function [28].

Heart rate (HR) is the speed of the heartbeat measured by the number of poundings of the heart per unit of time — typically beats per minute (bpm).

Respiration is the biochemical process in which the cells of an organism obtain energy by combining oxygen and glucose, resulting in the release of carbon dioxide.

B. Data Analysis

There is a set of multi-parameter data instances from healthy volunteers in The Stress Recognition in Automobile Drivers dataset [16] from the PhysioBank database [17], these data were taken while the volunteers were driving on a designed path including highways and city streets in the region of Boston. The aim of this work is to find out the feasibility of automated recognition of stress on the basis of the recorded signals, which include electrocardiogram (ECG), electromyography (EMG), galvanic skin resistance (GSR) measured on the hand and foot, heart rate (HR) and respiration.

We shall not use the entire dataset of seventeen drivers to study [16, 17] the stress degree, but to test the causal rate in the short and long term between predefined signals of these drivers. The general placement of sensors in automotive system is shown in Figure 1.

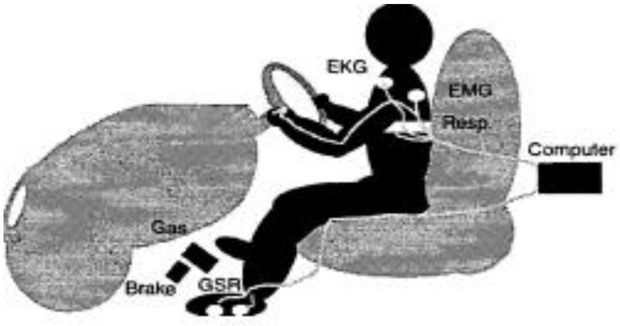


Fig. 1. Placement of sensors [29]

The final duration of the drive, with rest periods, varied from approximately 50 min to 1.5 h determined by traffic conditions. Drivers are questioned directly after each drive for filling out the subjective rating questionnaires. We placed the EMG on the trapezius muscle to indicate emotional stress [16] we measured the skin conductance in two places: in the middle and first finger of left hand with electrodes and on the left foot sole to measure the respiration we the expansion of chest cavity with an apropried sensor The EKG was sampled at 496 Hz, the skin conductivity and respiration sensor were sampled at 31 Hz, and the EMG was sampled at 15.5 Hz after first passing through a 0.5 s averaging filter. The signals were collected by an embedded computer in a modified car. The experimenter visually monitored the physiological signals as they were collected using a laptop PC running a remote display program. Figure 2 shows an example of the signals collected on a typical day's drive along with markings showing driving periods and events [29].

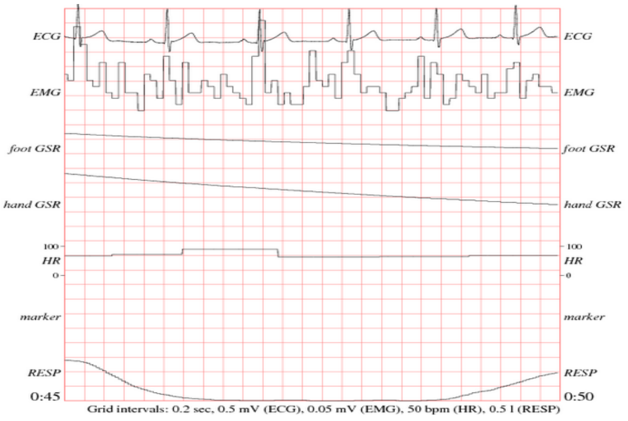


Fig. 2. Physiological data collected from Electrocardiogram (ECG), electromyogram (EMG), the respiration, heart rate, GSR foot and hand [16]

C. Methodology

Admittedly, the data models of panel have multiple advantages, but they do not seem sufficient any more to study all the phenomena, especially for our case resides in the study of the physiological signals. We must thus consider the last evolutions of the data of panel in term of multi-variety and non-stationariness in order to estimate our results correctly.

There exist a certain number of nonlinear models for data of panel, among which one can quote:

- Pooled Models.
- Fixed effect models.
  - ✓ Fixed effects estimation Models.
  - ✓ Existence of fixed effects Tests.
- Random effect models.
  - ✓ Estimation of the models for random purposes.
  - ✓ Hausmann Tests.
- Probit and Logit.
- Tobit I and II.
- Panel Co-integration data.

In what follows, one will be interested in this last method in order to evaluate our contributions. We will restrict our study with the models of Co-integration, of data, of panel, with the estimators FM-OLS and DOLS like with the causality of Granger in the panel.

To carry out, the objective is set higher; we called on a methodological strategy pluri-methodology, an analysis by the method of the panel data, which allowed us to exploit dimensions individual and temporal. Our approach of analysis is in the following stages:

- Unit root tests.
- Panel Co-integration,
- FM-OLS and DOLS Estimators,
- Panel Granger Causality.

Before starting the empirical part, it is necessary to explain each approach.

III. Co-Integration Approach

There exist a certain number of tests for Co-integration as a panel. One can quote, Kao [37], Bai and Ng [38], Mackoskey and Kao [39], Westerlund [40, 41, 42, 43], Westerlund and Edgerton [44], Hank [45, 46], Gengenbach, Palm and Urbain [47], Gutierrez [48] as well as the tests of Pedroni [21, 22, 23].

In our study and taking into account the length of the important temporal dimension of the data, we chose to apply the approaches of Pedroni.

Pedroni [21, 22] proposed various tests aiming at apprehending the worthless assumption of absence of intra-individual Co-integration at the same time for homogeneous and heterogeneous panels. Breaking values appearing in this work being relating to the presence of only one regressor in the relations of Co-integration.

Pedroni proposes an extension if the relations of Co-integration understand more than two variables. It is starting from these last tests that one concentrates more, because our study takes into account several parameters at the same time for the model as a panel. The tests of Pedroni take into account heterogeneity by the means of parameters which can differ between the individuals. Thus, under the alternative assumption, there exists a relation of Co-integration for each individual, and the parameters of this relation of Co-integration are not necessarily the same ones for each individual of the panel. The taking into account of such a heterogeneity constitutes an undeniable advantage since in practice, it is rare that the vectors of Co-integration are identical of one individual to the other of the panel.

Pedroni suggests seven tests: four are based on dimension will intra individual and three on inter individual dimension. In these seven tests, the statistics are built on the basis as of residues of the relations of Co-integration and a certain number of estimators of parameters of nuisance. By way of an example, the parameter of nuisance corresponds to the conditional variance of long run individual of the residues. Let us note finally that the number of delays retained in the regressions of type ADF (Increased Dickey-Fuller) can vary between the individuals. In order to implement the various tests, Pedroni suggests a procedure in four stages:

- Stage 1: One estimates the relation of long run and one recovers the estimated residues.
- Stage 2: For each individual, one differentiates the series yit and one calculates the residues resulting from the following regression.
- Stage 3: The variance of long run is estimated.
- Stage 4: By using the estimated residues, one chooses the suitable regression.

IV. Estimating the Long Run Relationship

The assertion of the existence of a relation of Co-integration between the series must be followed, by the estimate of the relation of long run. Several techniques exist in this direction, Pedroni [21, 22] showed that the two methods FM-OLS (Fully Modified Ordinary Least Squares) and also the method of least squares dynamic DOLS (Dynamic Ordinary Least Squares), are the adequate methods with the evaluation of such a relation and who allow the convergent estimate of the parameters in a panel presenting a problem of endogeneity and non stationariness.

A. FM-OLS

This procedure makes it possible to take account of the problems of endogeneity of the second order of the regressors (generated by the correlation between the residue of Co-integration and the innovations of variables I (1) present in the relation of Co-integration) and of the properties of autocorrelation and heteroscedasticity of the residues. It is in addition advisable to note that this model of estimate of the relation of long run is used much if we have a significant number of data, such physiological signals in the medical field.

B. DOLS

The approach DOLS was initially suggested in the case of the time series, then adapted to the case of the data of panel.

This technique consists in including values advanced and delayed in the relation of Co-integration, in order to eliminate the correlation between the explanatory variables and the term from error. The estimator DOLS has the same asymptotic distribution as estimator FM-OLS. This last show also a light superiority compared to the method DOLS. He is regarded as being the most robust technique in the estimate of the relations of panel Co-integration in the case of one a large number of data for the test. The representation of these two methods is illustrated in [49].

V. Panel Granger

As we saw in the previous section, Co-integration in data of panel is a method which makes it possible to check the existence or the absence of the long-term relation between the variables. It does not specify the direction of causality. When a relation of Co-integration exists between the variables, it must be modelled in a model with correction of dynamic error. The principal goal of each study is to draw up causal links between the endogenous variable and the whole of the exogenous variables. The tests of causality of Granger are based on the following regressions:

$$(1-L)\begin{bmatrix} X1_{it} \\ X2_{it} \\ X3_{it} \\ X4_{it} \\ \vdots_{it} \\ Xn_{it} \end{bmatrix} = \begin{bmatrix} a_{i\,X1} \\ a_{i\,X2} \\ a_{i\,X3} \\ a_{i\,X4} \\ \vdots \\ a_{i\,Xn} \end{bmatrix} + \sum_{i=1}^{P}(1-L)\begin{bmatrix} \vartheta_{11ip} & \vartheta_{12ip} \\ \vartheta_{21ip} & \vartheta_{22ip} \end{bmatrix}\begin{bmatrix} X1\ _{it-p} \\ X2\ _{it-p} \\ X3\ _{it-p} \\ X4\ _{it-p} \\ \vdots \\ Xn_{it-p} \end{bmatrix} + \begin{bmatrix} \beta_{X1\,i} \\ \beta_{X2\,i} \\ \beta_{X3\,i} \\ \beta_{X4\,i} \\ \vdots \\ \beta_{Xn_i} \end{bmatrix} Y\ _{t-1} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

Where $X1$, $X2$, … Xn represent the explanatory variables, p Lag (the length of delay), $(1-L)$ is the first operator of difference and Y t−1 signify the shifted term in correction of error coming from the relation of Co-integration. In order to illustrate all that, one proceeds to the empirical part relating to the theoretical aspect. One is interested first of all in the general information on the specification of the model used, concerning the various variables treated as well as the various panels of our study. Then, one follows in this part, the stages of the method of Co-integration in data of panel, while starting with the tests of the unit roots and while finishing by the test of causality of Granger in panel, all this for the three panels considered in our case.

The analysis of the biomedical signals nowadays is of an importance increased in the development of the medical therapeutic strategies. With the development of data processing and digital calculation, it becomes interesting to integrate an approach of assistance into the diagnosis in a computing process automatic. The choice of the mathematical models resulting from the signals for the characterization from such or such pathology becomes crucial then. We saw that in the previous chapter or decision making depends primarily on a certain number of parameters, difficult sometimes to extract.

In order to widen the detection of these anomalies, another vision is elaborate in this second contribution. It corresponds to the statistical analysis of observations regularly spaced in the time and in search of factors of causality between the physiological signals. More especially as these signals, generated by the displacement of an electric field in living tissue (EEG, EMG, ECG, etc), have multiple variations carrying relevant or harmful information to the extraction of medical information.

The criteria most used by the medical community are the measurement of time intervals (lasted of an event, separation of two events, delays) to characterize a temporal variation. This is why, we made a second contribution to develop a multivariate causal model between the cardiorespiratory myogalvanic signals. The end worked, after the distinction of the key factors which interest us to establish our goal, is the formulation of a robust statistical model which is a congruent representation of a stochastic process (unknown). In the world of the statistics/probabilities, there exist several models in this direction. Our approach in this article uses the models of causality within the meaning of Granger between the cardiorespiratory myogalvanic signals.

The causality concept represents the whole of the antecedents, which its intervention makes it possible to understand any phenomenon. In Mathematics, causality between two time series is generally studied in terms of the forecast improvement according to the characterization of Granger, or in aiming to impulsional analysis, according to the Sims principles.

In accordance with Granger sense, a series “causes” another series if the knowledge of the past of the first improves the forecast of the second. According to Sims, a series can be recognized like causal for another series, if the innovations of the first contribute to the variance of error of forecast of the second. Between these two principal modes of causality statistical characterization, the Granger approach is certainly that which had the most echoes among the mathematicians; it will thus be retained within the framework of this study. The base of the Granger definition is the dynamic relation between the variables. As indicated, it is stated in terms of improvement of the variable predictibility. For Granger and Sekkat, we cannot highlight causality without taking into account the “time” factor [19].

In our article, we begin the method of panel Co-integration in order to test the existence or absence of a long-term relation between our studied signals (cardiorespiratory electromyogalvanics signals) for the 17 drivers.

Its weak point is that it does not indicate the direction of this causality; hence, the necessity of tackling the so-called Panel Granger causal mathematical in order to model the directionality of the causality, the latter must be modeled in A dynamic error correction model of Engle and Granger [4].

Granger causality is one of the methods to model the idea of who causes the other, in other words, the idea that past effects help predict future effects. This concept was first traced by Wiener and implemented by Granger [4, 30, 32] as a linear autoregressive vector model VAR, and later generalized by John Geweke [31].

The main purpose of each study is to establish causal links between the endogenous variable and all exogenous variables, Granger causality tests will be based on the following regression:

$$(1-L)\begin{bmatrix} ECG_{it} \\ EMG_{it} \\ FOOT\ GSR_{it} \\ HAND\ GSR_{it} \\ HR_{it} \\ respiration_{it} \end{bmatrix} = \begin{bmatrix} a_{i\,ECG} \\ a_{i\,EMG} \\ a_{i\,FOOT\ GSR} \\ a_{i\,HAND\ GSR} \\ a_{i\,HR} \\ a_{i\,RESPIRATION} \end{bmatrix} + \sum_{i=1}^{P}(1-L)\begin{bmatrix} \vartheta_{11ip} & \vartheta_{12ip} \\ \vartheta_{21ip} & \vartheta_{22ip} \end{bmatrix}\begin{bmatrix} ECG_{\ it-p} \\ EMG_{\ it-p} \\ FOOT\ GSR_{\ it-p} \\ HAND\ GSR_{\ it-p} \\ HR_{\ it-p} \\ respiration_{it-p} \end{bmatrix} + \begin{bmatrix} \beta_{ECG\ i} \\ \beta_{EMG\ i} \\ \beta_{FOOT\ GSR\ i} \\ \beta_{HAND\ GSR\ i} \\ \beta_{HR\ i} \\ \beta_{RESPIRATION_i} \end{bmatrix} ECT_{\ t-1} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

p is the lag (the delay length), L-1 is the first difference operator and ECT_{t-1} means the offset error correction term from the relationship of the Co-integration. An error correction model can distinguish between long-term and short-term relationship in the Granger causality. The short-term dynamics are captured by different coefficients of staggered terms.

The statistical significance of the coefficients of each explanatory variable is used to test Granger causality in the long term while the meaning of the coefficients of ECT_{t-1} provides information on the short-term causality.

VI. Empirical Result

Recent progress of mathematical modelling filled with enthusiasm the researchers, in particular in the panel data analysis. Methodology that we apply here is mainly based on the four fundamental parts of the panel data analysis, exposed previously. Initially, we applied unit roots tests to evaluate the series stationnarity.

The second phase consists to apply the Co-integration tests and to quantify later this long-term relation using FM-OLS and DOLS estimators. Finally, we applied the Granger causality tests to the whole of the studied panels.

Before beginning our results, we must clarify some details and signs:

- ⇄ Corresponds to the causal direction between our physiological signals.
- ▬ this sign is a non-causal
- A Value above the sign ⇒, is the value of F-statistic, which is considered a measure of the correlation between the variables studied.
- A Value below the sign ⇒, which is in brackets, corresponds to the value of the probability of causation.
- Level: I (0).
- First difference: I (1).
- {x} : Std Error (The standard error is the standard deviation of the sampling distribution of a statistic)
- *: Indicates statistical significance at 1%.
- x E y : (*x* exponential *y*).
- [ ]: Long-term coefficient.

A. Panel unit root tests

In this section, we used the unit root tests on panel data (Levin Lin and Chu (LLC) [18], IM Pesaran and Shin (IPS) [19], Breitung (BRT), Maddala and Wu (MW) [20]). We will introduce you now; the results of LLC, IPS, BRT, MW and HADRI tests applied to the variables of our model. Our analysis of the results is mainly based on the Hadri test, which is known and recognized by its robustness, power and precision, it shows us without any ambiguity that the variables are non-stationary in level. Although the results of other tests used can be confusing given that reveals a level stationarity, this incompatibility is due to the statistical differences of the various tests used that can give distinct results. This divergence in the results and power of HADRI test requires us to test and verify the stationarity of our variable in first differences.

B. Panel Co-integration

After the checking of the non-stationarity for the all variables of the panel, we proceed to study the existence of one long-term relation between these variables, and this by applying the Pedroni Co-integration tests, which are based on the unit roots tests of estimated residues, trying now to test Co-integration for the signals. Pedroni [21, 22, 23] proposes two tests families, one realized in 1999 [21] resting on seven tests (four based on intra individual dimension and three on inter-individual dimension) and another family of tests realized in 2004 [22], suggesting another four tests containing balanced statistics. The two categories of tests rest on the null hypothesis of absence of Co-integration. The co-integration of the variables depends on the value of the probability associated with each statistics (probability <0.01). Table.1 summarizes the results of the Pedroni Co-integration statistics. From the results of the Pedroni Co-integration tests, we can notice that among the whole of the statistical tests, all the probability values are less than 1% (they all are to 0.0000). Therefore, the whole of these tests shows the existence of a relation of Co-integration.

TABLE I. PEDRONI CO-INTEGRATION FOR ECG

Pedroni Methods	Within dimension			Between dimension		
	Test	**Statistics**	**Prob**	**Test**	**Statistics**	**Prob**
Pedroni (1999) [21]	**Panel v-statistic**	27.52606	(0.000)*	**Group ρ-statistic**	-43.57195	(0.0000)*
	Panel rho-statistic	-48.9354	(0.000)*	**Group pp-statistic**	-24.29321	(0.0000)*
	Panel PP-statistic	-24.7533	(0.000)*	**Group ADF-statistic**	-68.54084	(0.0000)*
	Panel ADF-statistic	-46.7641	(0.000)*			
Pedroni (2004) (Weighted statistic) [22]	**Panel v-statistic**	18.80569	(0.000)*			
	Panel rho-statistic	-45.0907	(0.000)*			
	Panel PP-statistic	-23.7351	(0.000)*			
	Panel ADF-statistic	-46.4090	(0.000)*			

C. *FM-OLS and DOLS estimations*

In the light of the projections carried out in non-stationary time series, the estimators of non-stationary panel data can still solve a certain number of problems, in particular on the level of the estimate and inference. To estimate Co-integrated variables systems, just like to carry out tests on the Co-integration vectors, it is necessary to use an effective estimate methods. FM-OLS and DOLS estimators proposed by Pedroni Kao and Chiang [24] and Mark and Sul [25].

The estimation of a Co-integration relation, if it exists, which connects the variables of the model in double index is established by the suitable method (FM-OLS and/or DOLS). The estimated parameters by one of these methods will be interpreted as being long run elasticities. It is important to emphasize that the DOLS method presents the disadvantage of reducing the number of freedom degrees of the studied variables, which leads to less reliable estimates. The estimation results are reported in the Table.

Table.2 establishes long-term elasticity between variables of the model using FM-OLS and DOLS estimators. The modelling of within dimension enables us to take into account the heterogeneity of the coefficients in their temporal and/or individual dimension. The within estimator eliminates the individual specific effects.

By analyzing the estimated model by the FM-OLS regressor, The results for ECG-dependent suggests that an increase of 1% in EMG, FOOTGSR, HAND GSR, HR, and RESPIRATION implies a variation of ECG which take a value of 0.016248%, 0.007241%, 0.028366%, 0.000511% and 0.000110% respectively in the within dimension based on the method FMOLS. With same way, the result for ECG suggest that an increase of 1% in EMG, FOOT GSR, HAND GSR, HR and RESPIRATION implies a variation of ECG which take a value of 0.065684%, 0.014534%, 0.032800%, 0.000304%, 0.005986% respectively in the between dimension based on the same method.

TABLE II. FM-OLSS AND DOLS FOR ECG

Dependent Variable « ECG »	FM-OLS				
Independent variables	**EMG**	**FOOT GSR**	**HAND GSR**	**HR**	**RESPIRATION**
Within Results	[-0.016248] {0.003222} -5.042276 (0.0000)*	[0.007241] {0.007033} 1.029656 (0.3032)	[-0.028366] {0.005027} -5.643077 (0.0000)*	[0.000511] {0.000387} 1.321070 (0.1865)	[-0.000110] {0.001251} -0.088084 (0.9298)
Between Results	[0.065684] {0.017788} 3.692641 (0.0002)*	[0.014534] {0.094826} 0.153274 (0.8782)	[0.032800] {0.168461} 0.194703 (0.8456)	[0.000304] {0.000402} 0.755944 (0.4497)	[-0.005986] {0.002573} -2.326718 (0.0200)
Dependent Variable « ECG »	DOLS				
Independent variables	**EMG**	**FOOT GSR**	**HAND GSR**	**HR**	**RESPIRATION**
Within Results	[-0.008268] {0.001557} -5.309399 (0.0000)*	[0.011119] {0.003388} 3.282318 (0.0010)*	[-0.028238] {0.002420} -11.66741 (0.0000)*	[0.000366] {0.000187} 1.960664 (0.0499)	[-0.000348] {0.000602} -0.578059 (0.5632)
Between Results	[-0.008414] {0.001553} -5.417621 (0.0000)*	[0.011257] {0.003385} 3.325155 (0.0009)*	[-0.028268] {0.002422} -11.67275 (0.0000)*	[0.000373] {0.000187} 2.001192 (0.0454)	[-0.000363] {0.000602} -0.603074 (0.5465)

D. Panel Granger causality

The purpose of this part is to test the causal links between these variables using the Panel Granger causality test. A Granger causality analysis is carried out in order to determine if there is a power of potential foreseeability from one indicator to another. The results of the test for the all variables are summarized in table.3. It should be noted that the optimal delay (Lag) was established using the Akaike and Schwarz information criterion [50, 51].

The purpose of our study is to show the interactive relations between the whole of the signals, but that does not preclude under investigation of all possible relations. From the causality tests results presented in table.3, we can deduce the causal links direction which may appear between the variables in the threshold criticizes (probability of error) of 1%. To be more explicit, if the probability is less than 1%, we speak about a causal relation, in the opposite case, we speak about a no causal relation between variables.

TABLE III. PANEL GRANGER CAUSALITY

Lag=49	ECG	EMG	FOOT GSR	HAND GSR	HR	RESPIRATION
ECG	✗	1.01343 ▭ (0.4462)	3.76621 ← (2.E-13)*	3.76640 ← (2.E-13)*	0.93217 ▭ (0.5853)	3.17198 ← (5.E-10)*
EMG	2.08195 → (0.0001)*	✗	0.23925 ▭ (1.0000)	0.22863 ▭ (1.0000)	0.07047 ▭ (1.0000)	0.34500 ▭ (0.9999)
FOOT GSR	6.73692 → (2.E-32)*	0.30623 ▭ (1.0000)	✗	446.983 ← (0.0000)*	0.08348 ▭ (1.0000)	943.990 ← (0.0000)*
HAND GSR	6.82513 → (5.E-33)*	0.05842 ▭ (1.0000)	450.845 → (0.0000)*	✗	0.46971 ▭ (0.9972)	1002.34 ← (0.0000)*
HR	0.80450 ▭ (0.7912)	0.16269 ▭ (1.0000)	1.95451 → (0.0005)*	1.85638 → (0.0014)*	✗	2.43276 ← (3.E-06)*
RESP	5.61495 → (5.E-25)*	0.10165 ▭ (1.0000)	1.19212 ▭ (0.1989)	15.5611 → (3.E-95)*	1.09573 ▭ (0.3185)	✗

Our study aims to illustrate the interactive relationships between all the variables EMG, GSR FOOT, HAND GSR, HR, RESPIRATION and the ECG signal, but that does not preclude the study of all possible relationships.

From the results of Granger Causality Test Panel presented in the table above, we can deduce the direction of causal relationships between variables can figure the critical threshold (error probability) of 1%.

VII. DISCUSSION AND CONCLUSION

In order to check the long-term convergence between our studied signals, we applied the Co-integration method; the results show that there is actually a convergence of these signals. In this article, we focus on the technical and non-medical aspects in the fact that we belong to the telemedicine fields and we try to trace the short-and long-term causalities among drivers to take the necessary statistical information to prevent damage accidents. The knowledge and the quantitative understanding of these interactions are critical in monitoring people during driving, we wanted to study causality tests while driving in order to develop in next research an application telemedicine preventive, with perfect causal analysis of vital signs during driving.

These results and these statistical analyses will constitute at the same time a prediction base as well as a beginning of action towards new researches orientations treating the physiological interactions quantitatively. In this article, we concentrated much more on the technical sides and we leave the medical explanations to the health specialists to clarify this convergence/divergence, causality/no causality of the physiological signals.

The fact of studying these interactions, perhaps we can cure and prevent the sudden death which is quasi-unforeseeable and relentless, implying sudden demonstrations of some undesirable interactions in the human body, consequently it became a true syndrome in most unsuspected cases for the majority of the population (Nutritive, adult,…).

As prospects with these research tasks, we suggest the integration of several algorithms of the signal treatment, such as the causal processes, on a system embarked to appreciate these interdependences between physiological signals in the same application. The obtained results in this article pointed out the importance of improving the models existing in order to a better description for the various phenomena. However, in order to understand the mechanisms and to be able to prevent and treat more effectively these diseases, our studied models require to be improved by integration of the qualitative factors like the age, the sex, the diseases histories … and that due to a certain number of mathematical models like Logit, Probit and Tobit.

We also plan to work out a model embarked to study panels more widened and with a very large number of patients gathered in blocks having the same symptoms, or the same age interval, or the same kind…

CONFLICT OF INTEREST

The authors confirm that this article content have no conflict of interest.

REFERENCES

[1] E. Bekiaris, S. Nikolaou, "State of the Art on Driver Hypovigilance Monitoring and Warning Systems, AWAKE System for effective Assessment of driver vigilance and Warning According to traffic risk Estimation, " (IST-2000-28062). 2001 Nov; 20.

[2] Accidents de la route: 4540 morts et 69.582 blessés en Algérie en 2013 Available at: http://www.reflexiondz.net/ACCIDENTS-DE-LA-ROUTE-4540-morts-et-69-582-blesses-en-Algerie-en-2013_a28409.html

[3] J. A. Horne, Reyner. L. A, "Counteracting driver sleepiness: effects of napping, caffeine and placebo. Psychophysiology," 33 (3): 306-309; 1996 May, doi: 10.1111/j.1469-8986.1996.tb00428.x.

[4] Granger. C. W. J, "Investigating causal relations by econometric and cross-spectralmethods, " Econometrica, 1969 Aug; 37 (3), 424–438.

[5] Yongmiao. Hong, Yanhui. Liu, Shouyang. Wang, "Granger causality in risk and detection of extreme risk spillover between financial markets, " Journal of Econometrics, 2009 june; 150 (2), 271–287.

[6] Pierre. Olivier. Amblard, Olivier. J. J. Michel, "On directed information theory and Granger causality graphs, " J ComputNeurosci, 2011, 30,pp. 7–16.

[7] Xiang. Li, Kaiming. Li, Lei. Guo, Chulwoo. Lim, "Tianming Liu. Fiber-Centered Granger Causality Analysis. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011," Lecture Notes in Computer Science. 2011, 6892,pp. 251-259. doi: 10.1007/978-3-642-23629-7_31.

[8] Zhu. J, Chen. Y, Leonardson. A. S, Wang. K, Lamb, J. R, et al, "Characterizing dynamic changes in the human blood transcriptional network," PLoSComput. Biol. 2010, 6 (2) , e1000671. doi:10.1371/journal.pcbi.1000671.

[9] Pereda. E, R. Q. Quiroga, J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," ProgNeurobiol. 2005,77 (1-2),pp. 1–37.

[10] Yifan. Zhao, Steve. A. Billings, Hua. Liang. Wei, Ptolemaios. G. Sarrigiannis, "A parametric method to measure time-varying linear and nonlinear causality with Applications to EEG data," IEEE, 2013, pp.1-7.

[11] M. Palu. S, A. Stefanovska. Phys. Rev. 2003; E 67, 055201R.

[12] Faes. L, L. Widesott, M. Del. Greco, R. Antolini, G. Nollo, "Causal cross-spectral analysis of heart rate and blood pressure variability for describing the impairment of the cardiovascular control in neurally mediated syncope," IEEE Trans Biomed Eng, 2006, 53 (1),pp. 65–73.

[13] Samir. Ghouali, Mohammed. Feham,Yassine. Zakarya. Ghouali, "The direction of information between Cardiorespiratory Hemodynamic signals: Test Analysis using Granger Causality," Journal of Mathematics, Statistics and Operations Research (JMSOR), 2(2), DOI: 10.5176/2251-3388_2.2.52. doi: 10.5176/2251-3388_2.2.52.

[14] Pereda. E , D. M. de. La. Cruz, L. De. Vera, J. J. Gonzalez, "Comparing generalized and phase synchronization in cardiovascular and cardiorespiratory signals," IEEE Trans Biomed Eng. 2005; 52 (4): 578–583.

[15] Rosenblum, M. G, L. Cimponeriu, A. Bezerianos, A. Patzak, R. Mrowka, "Identification of coupling direction: application to cardiorespiratory interaction," Phys. 2002; Rev. E 65:041909.

[16] PHYSIOBANK ATM Available at: http://www.physionet.org/cgi-bin/atm/ATM.

[17] PhysioBank: Physiologic signal archives for biomedical research Available at: http://www.physiomet.org/physiobank.

[18] Levin. A, Lin. CF, Lin. ChuJ, "Unit root tests in panel data asymptotic and finite-sample properties," J Econometrics. 2002, 108 (1),pp. 1–24.

[19] Im. KS, Pesaran. MH, Shin. Y, "Testing for unit roots in heterogeneous panels," J of Econometrics, 2003, 115,pp. 53–74.

[20] Maddala. G. S, Wu. S, "A comparative study of unit root tests with panel data and a new simple test," Oxford Bulletin of Economics and Statistics Special Issue, 1999, 61, pp.631–652.

[21] Pedroni P, "Critical values for Co-integration tests in heterogeneous panels with multiple regressors. Oxford Bulletin of Economics and Statistics," 1999, 61, pp.653–678.

[22] Pedroni. P, "Panel Co-integration: asymptotic and finite sample properties of fooled time series tests with an application to the PPP hypothesis," Econometric Theory , 2004, 20,pp. 597–625.

[23] Pedroni. P, "Purchasing power parity tests in Co-integrated panels. The Review of Economics and Statistics," 2001 November, 83(4),pp. 727–731.

[24] Kao. C, Chiang. M. H, "On the estimation and inference of a Co-integrated regression in panel data," In: Baltagi, B.H. (Ed.), "Advances in Econometrics: Nonstationary Panels. Panel Co-integration and Dynamic Panels," 2000, 15, pp.179–222.

[25] Mark. N. C, Sul. D, "Co-integration vector estimation by panel DOLS and long-run money demand.Oxford Bulletin of Economics and Statistics," 2003 Dec, 65 (5),pp. 665-680.

[26] Nagal. D,Sharma. S,"Simultaneous 12-lead QRS detection by K-means clustering algorithm," Recent Advances and Innovations in Engineering (ICRAIE), 2014, pp.1 - 4, DOI:10.1109/ICRAIE.2014.6909244

[27] I. Elamvazuthi,N.H.X. Duy,Zulfiqar Ali,S.W. Su,M.K.A. Ahamed Khan,S. Parasuraman ,"Electromyography (EMG) based Classification of NeuromuscularDisorders using Multi-Layer Perceptron," Procedia Computer Science, 2015 IEEE International Symposium on Robotics and Intelligent Sensors (IEEE IRIS2015), Volume 76, 2015, Pages 223-228.

[28] Nicola. Gerrett,Bernard. Redortier,Thomas. Voelcker, George. Havenith, "A comparison of galvanic skin conductance and skin wettedness as indicators of thermal discomfort during moderate and high metabolic rates," Journal of Thermal Biology, Volume 38, Issue 8, December 2013, Pages 530–538, http://dx.doi.org/10.1016/j.jtherbio.2013.09.003.

[29] Jennifer. A. Healey, Rosalind. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," IEEE Trans. Intelligent Transportations Systems, Vol. 6, No. 2, April 2005.

[30] Anil. K. Seth, Adam. B. Barrett, Lionel. Barnett,"Granger Causality Analysis in Neuroscience and Neuroimaging," The Journal of Neuroscience, February 25, 2015, 35(8):pp. 3293–3297.

[31] Geweke. J, "Measurement of linear dependence and feedback between multiple time series," J Am Statistical Assoc, 77, pp.304 –313,1982

[32] Barnett. L, Barrett. AB, Seth. AK, "Granger causality and transfer entropy are equivalent for Gaussian variables," Phys Rev Lett 103, 238701, 2009.

[33] Samir. Ghouali, Mohammed. Feham, Yassine. Zakarya. Ghouali, "Causal relationships between Cardiorespiratory Hemodynamics signals: Test Analysis using panel cointegration," The World Congress On Computer Applications and Information Systems (WCCAIS'2014), 2014 January 17-19, Tunisia, IEEE, pp 1-8. DOI: 10.1109/WCCAIS.2014.6916591.

[34] Samir. Ghouali, Mohammed. Feham, Yassine. Zakarya. Ghouali, "Revealing the Dynamic Correlation between Cardiac and Respiratory Hemodynamic Signals Using Time-Dependent Panel Co-Integration Analysis,". doi: 10.15662/ijareeie.2014.0311091, November 2014.

[35] Sami Khedhiri, "cours d'économétrie méthodes et applications", Lavoisier 2007 paris page 93.

[36] Hsiao, C, "Analysis of panel data", Cambridge University Press, 2003.

[37] Kao C, "Spurious Regression and Residual-Based Tests for Co-integrationin Panel Data", 1999, Journal of Econometrics, 90, pp. 1-44.

[38] Bai J, Ng S, "A panic Attack on Unit Roots and Cointegration", 2004, Econometrica, 72(4), pp. 1127-1178.

[39] McCoskey S. et Kao C, "A Residual-Based Test of the Null of Co-integrationin Panel Data", 1998, Econometric Reviews, 17, pp. 57-84.

[40] Westerlund, J, "Testing for error correction in panel data", 2007, Oxford Bulletin of Economics and Statistics 69: 709–748.

[41] J. Westerlund, "New simple tests for panel Cointegration", Econometric Reviews, 24:297–316, 2005.

[42] J. Westerlund, "Testing for panel Co-integrationwith a level break". Economics Letters, 91:27–33, 2006.

[43] J. Westerlund, "Panel Co-integrationtests of the Fisher Hypothesis", Journal of Applied Econometrics, 23:193–233, 2008.

[44] J. Westerlund and D. L. Edgerton, "Simple tests for Co-integrationin dependent panels with structural breaks", Lund University, Department of Economics, January 2007.

[45] C. Hanck, "Cross-sectional correlation robust tests for panel Cointegration", Universität Dortmund, SFB 475 Technical Report 44/06, Ruhr Graduate School in Economics, November 2006.

[46] C. Hanck,"A meta analytic approach to testing for panel Cointegration", Universität Dortmund, SFB 475 Technical Report 02/07, Ruhr Graduate School in Economics, January 2007.

[47] Christian Gengenbach, Franz C. Palm, Jean-Pierre Urbain, "Co-integrationTesting in Panels with Common Factors", Oxford Bulletin of Economics and Statistics, Volume 68, Issue s1, December 2006 Pages 683–719, DOI: 10.1111/j.1468-0084.2006.00452.x.

[48] L. Gutierrez, "Simple tests for Co-integrationin panels with structural breaks", Applied Economics Letters, 2008.

[49] Roberto Basile , Mauro Costantini , Sergio Destefanis, "Unit root and cointegration tests for cross-sectionally correlated panels. Estimating regional production functions," ISAE Working Paper No. 53. Available at SSRN: https://ssrn.com/abstract=936324 or http://dx.doi.org/10.2139/ssrn.936324

[50] Renaud Lacelot, Matthieu Lesnoff, "Sélection de modèles avec l'AIC et critères d'information dérivés," Version 3, Novembre 2005.

[51] Henry de-Graft Acquah, "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship," Journal of Development and Agricultural Economics Vol. 2(1) pp. 001-006, January 2010, Available online at http://www.academicjournals.org/JDAE ISSN 2006- 9774.

Internet of Things (IoT) : Charity Automation

Maher Omar Alshammari
College of Computer Sciences & IT
King Faisal University,
Saudi Arabia

Abdulmohsen A. Almulhem
College of Computer Sciences & IT
King Faisal University,
Saudi Arabia

Noor Zaman
College of Computer Sciences & IT
King Faisal University,
Saudi Arabia

***Abstract*—People are living in cities and villages based on their profession and their earnings. Those who have better earnings can live their live nicely. However, those who do not have good earnings are facing difficulties to survive their lives even for their basic necessities such as food and clothes. Government and limited charity organizations are trying to help them. In the Kingdom of Saudi Arabia few charity organizations placed few donation boxes around the city to collect donations on donor's ease, but it has become hard for them to monitor them regularly, which affect the donation condition. Involving the Internet of Technology (IoT) will give the donors' comfort and fast way to communicate with the charity, which will make the donation process efficient, easier and in well-organized as well. This paper presents a smart solution which is based on advanced technologies namely; Smart Charity (SC) that will help charity organizations, donators and needy people by involving IoT. SC working mechanism based on two folds, 1)Web-Based Application and 2)Android-based smart Application that will enable donors to donate through their mobiles anywhere and anytime, as well they can suggest the best time for them so the charity organization's representative can visit and collect the desired donations. SC will enable the charity organization to know the location of donors and needy people through GPS as well. In addition, SC introduces Smart Donation Box (SDB) concept by involving IoT, which will have the capability to communicate with charity organizations about its current status such as quarterly, half or fully filled.**

Keywords—Smart city; Smart Charity; Internet of Things (IoT)

I. Introduction

Willing to help needy people is a great nobleness. It is not hard to help other people, especially if the help process is simple and easy for you, and for others it could mean everything for them. That is donating which is an active way of helping others. In return, Smart Charity (SC) project produced a system, which aims for helping needy people (people who need money, clothes and food). One of the main goals for SC is to replace old fashion of paper work, and involve the Internet of Things (IoT) to help make better donation boxes that called Smart Donation Box (SDB). With a wide range of technologies available across our daily life, we have found an appropriate way by producing smart system that is the combination of an Android smart phone application and a Web-based application. Donors are able to use the application to enter their information and specify the location with the help of GPS coordinates. On the other hand, the web-based application enables charity organizations to manage the process of donation. Finally, the Smart Donation Boxes are able to notify/communicate with the charity organization timely to ensure that the donations are not wasted.

II. Background

In Saudi Arabia, people are willing to help and donate for needy people. Based on official statistics from the Alber Charity Organization in Al-Ahsa at 2015 that the number of physical donations is 12,787,060 SR [1]. In fact, the number is larger, but huge amounts of expected donations are wasted because there is no efficient way to collect and distribute them.

Currently there are several applications which exist to help in the donation process, such that Saudi Food Bank "Eta'am", "Makkah charity", " NemahKeep " and "Alber Charity".

According to our investigation and visiting to these organizations, the donators must come to the charity offices to make the donation, while with Eta'am charity you have to contact them and send an earlier request before you have any ceremony. As well with the other application NemahKeep, the donator must call the responsible for collecting the donations by cell phone and tell them the address. What about in the large cities it will be difficult for the charity to know the exact location. In big cities, it will be hard for charity organizations to keep track of all donation boxes and collect them timely. Dar_Alkhair recently launch new smart donation box. However, the main function of the new box is change light color from green to the red at the top of the box to intimate the donors will know that the box is filled.

Smart Charity project will expand the existing applications by providing the efficient search about the donator's location through (GPS) and the convenient time for collecting the donation. In addition, SC sends a notification message to the needy people if any new donation arrives for them, and what is the best time that the charity may send the donation to them. SC are stores the information of donators and needy people privately and securely in a database.

In the same time, SC will address big issue "Ignored donation boxes", that none of the existing applications takes good care of it. SC will provide an easy fast efficient solution for the charity donation boxes, by providing smart hardware device depend on IoT [2] that will forward a message to the charity organization whenever the donation box is filled.

III. Literature Review

This section provides a detailed literature review specifically about the existing applications for the donation, which are trying to increase and activate the donation process, and make it easier.

There are numbers of existing systems and charities organization that are increasingly trying to activate the donation process, to make it more suitable and use the technology to reduce the effort of the donator. This section will explore some charities donation process and recent works in this field one by one.

Eta'am charity organization [3] is a charity organization that mainly focuses on helping needy people by collecting extra healthy food from different parties and delivers it to needy people after packing. Their process of donation is that the donator should contact the charity and tell them about the type of the food and best suitable time and date to collect the food. Eta'am uses a website that shows the latest news, updates about the organization, and allows the donors to donate through their website as well. Eta'am also, uses a mobile application to do the same job as the website. However, following are the limitations of this system.

- Only accepting food donations from big ceremony minimum invited are 70 guests.
- The donator cannot create account, each time they have to fill up the form repeatedly and send it to the charity.
- There is no interaction between the donors and the charity.
- The needy people cannot register them self in easy way in order to receive donations.
- The registered needy people by the charity cannot update their location automatically in case if their location is changed.

Makkah charity organization [4] is a charity organization that helps the needy people by receiving donations at their offices on specific scheduled time or by bank transferring through their accounts. They have several programs one of them is helping needy people by providing them the necessary living hood. However, following are the limitations of this system.

- They use the website for showing their latest programs and their bank account details only.
- Donators have to visit the charity organization in order to donate physical donation like 'Clothes'.
- There is no interaction between the donors and the charity.
- The needy people cannot register them self in easy way in order to receive donations.
- The registered needy people by the charity cannot update their location automatically in case if their location is changed.
- The needy people cannot set a priority of their need.

NemahKeep [5] is helping by collecting unneeded healthy food from the ceremonies and hotel restaurants then organize it again to deliver it for Makkah visitors and needy people. Their process of donation is that the donator should contact the responsible of NemahKeep by phone and give them the location, the best time and date to collect the food. NemahKeep charity uses a Twitter account to show the latest news and other statics about the donations, as well they have a YouTube channel for showing some videos that motivate people to donate. NemahKeep uses the cellphone for communication with the donator. However, following are the limitations of this system.

- They do not use the technology in efficient way to help in donation process.
- Donators have to call responsible of NemahKeep in order to donate.
- There is no easy way that gives the donors motivation to donate.
- The needy people cannot register them self in easy way in order to receive donations.
- The registered needy people by the charity cannot update their location automatically in case if their location is changed.

Alber charity organization & Dar_ Alkhair charity organization [1] are one of the biggest charity organizations in Saudi Arabia that collect money from the government, companies, rich people and donators. As well, the charity collects the physical donations through donation boxes around the city. The charity has a website and a mobile application for showing the latest news about the charity and their work. However, following are the limitations of this system.

- They do not use the technology in efficient way.
- Donators have to visit the charity organization in order to donate physical donation like 'Clothes'.
- The charity ignores the donation boxes for long time.
- The cost of checking the donation boxes is high and makes a lot of effort to the charity employee.
- There is no mechanism to check the status of donation boxes.
- The needy people cannot set a priority of their need.

Rahman, Akter, Hossain, Basak and Ahmed [5] they proposed Smart Blood Query, which use smart mobile application in order to enable donator to use the app they have to register their necessary information including blood type and the current location, the app enable the blood seekers to see the nearest blood donators, and contact with them through SMS. In case if the blood donator's response with 'NO' a new search will be initiated to find the nearest blood banks and SMS will send to the recipient, provide them with some information, the blood donors have option to accept the donation request by reply to the SMS 'YES' or deny it by reply 'NO'. The App ass well, enables donator to update their information include location and availability. However, following are the limitations of this system.

- The system does not use GPS future in order to update the donor's location automatically.
- There is no confidentiality of the donor's information.
- The system is for blood donation only.

AlDossari, AlMubarak, AlBukhowa, and AlSaif [7] they proposed smart system namely: Life Doners, which is an application that try to address the issues related to blood donation in the emergency. Life Doners are trying to link the patients, donors and the hospital in one system, the proposed system have an android app which will be used by the donator to specify their information along with the suitable time for them, so in case if there is emergency situation required the donors blood the hospital can search for the nearest suitable donor. As well, the proposed system using a website for the hospital to manage the blood donation process effectively. However, following are the limitation of this system.

- The proposed system is for blood donation only.

Khan and Qureshi [8] the proposed system is a web-based 'blood management system' for Pakistanis citizens. In the proposed system the blood donator registers them self in the system and fill up the necessary information along with the contact number and blood type, the patients can contact blood donors through the system or by the phone in order to achieve the required blood donation. Overall, the proposed system is useful for the admin and patients to know the contact number of the blood donators. However, following are the limitation of this system.

- The proposed system is website only.
- Donors cannot specify their available time to donate blood.
- Does not use GPS features in order to know the location.
- There is no confidentiality of the donor's information.
- The proposed system is for blood donation only.

AlHashim, Al-Madani, Al-Amri, Al-Ghamdi, Bashamakh, and Aljojo [9] they proposed a Blood Management System that allows the willing donors to donate, and help the registered hospitals in the system to keep a record of the donor's data in case if they need to communicate with them. The proposed system easily enable the hospitals to register in the system and enter what type of blood they in need. The target of this system is to make it easier for donors to know which hospitals are in need for their blood type in case if they want to donate. However, following are the limitation of this system.

- The proposed system use website only.
- The proposed system is for blood donation only.
- Does not use GPS features in order to know the hospital location.

IV. SC System Analysis

This section explains how the Smart Charity system will work.

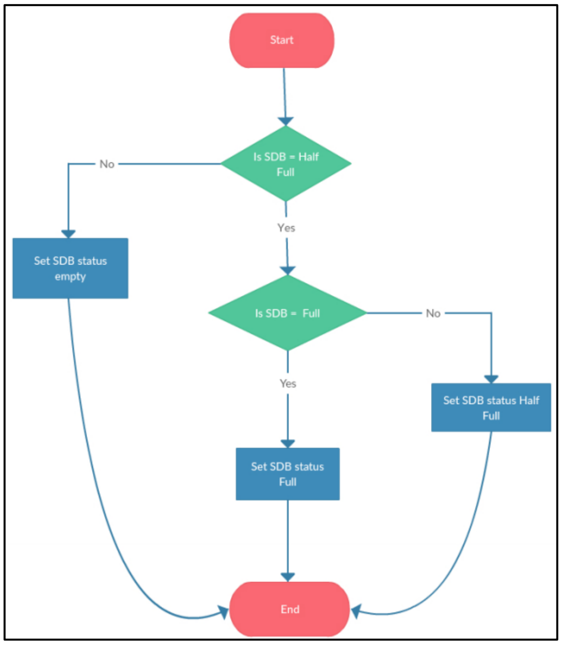


Fig. 1. Flow chart of smart donation box

Figure 1 shows the flow chart that explains the flow of the Smart Charity Donation box device, at the beginning the Arduino Uno start the connection with the charity organization website server and check the status of the box if the low-level sensor is cut and the high level are cut also that means the donation box is full, in this case, the system will update the status of the donation box to full. Otherwise, if only the low-level sensor is cut the donation box status will update it half-full and connection with the server close.

```
Step 1: Start Connection
Step 2: Check the low level sensor "half full"
Step 3: If low level sensor are cut
    Step 4: check the low level sensor "full"
    Step 5: If low level sensor are cut
            Display donation box are full
        Else
            Display donation box are Half-full
    Else
        Display donation box are empty
Step 6: Close Connection
```

Fig. 2. Smart Donation Box algorithm

Figure 2 shows the donation box algorithm, in the beginning, the Arduino start a connection with the charity server, after that the sensor scan and deduct the objects inside the donation box. In case if the items reach to a certain level the Arduino will notify the charity.

V. SC SYSTEM GUI DESIGN

This section shows the system GUI design which includes the smart box donation SBD prototype

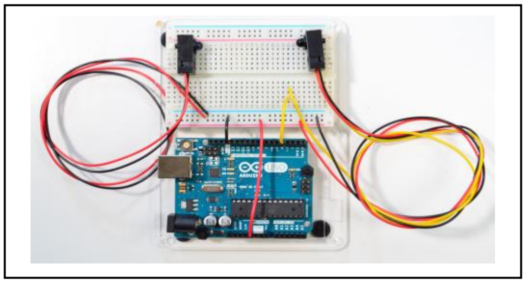


Fig. 3. Main component of the donation box

Figure 3 shows the main component of the Smart Donation Box Device, which is Arduino Uno board used as a microcontroller. Infrared IR Sensor used to determine the status of the donation box. Arduino Wi-Fi shield used for communications with the charity organization server.

Figures 4 shows the registration page for the donators as well for the needy people, needy people will have to check the checkbox and enter more information after the user(donator or needy people) complete the information and click on register button they will see a notification message in case of the registration completed or failed

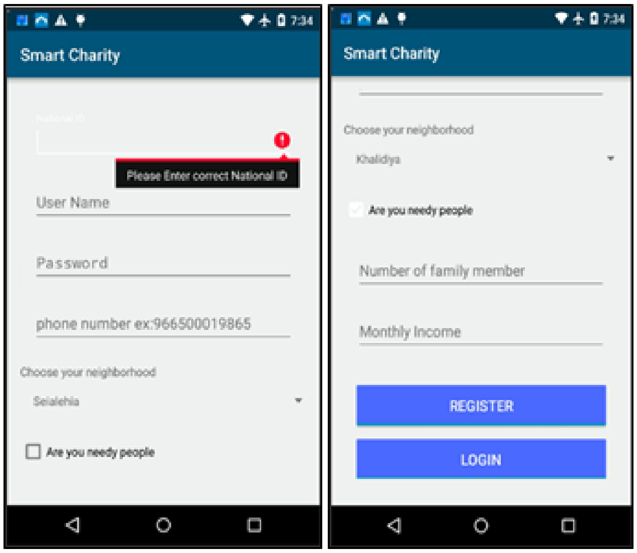


Fig. 4. Android app registration page

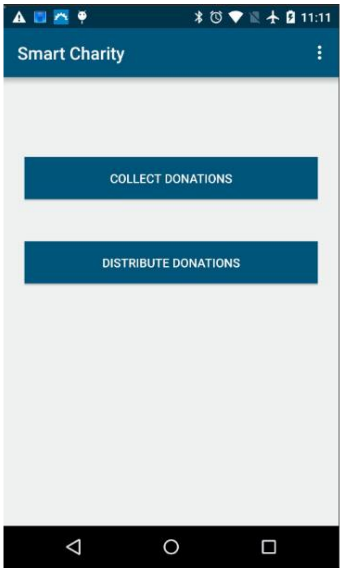
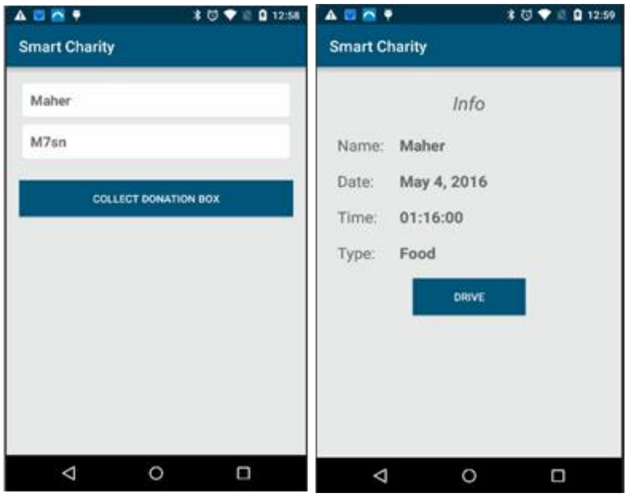


Fig. 5. Android app donator's page

Figure 5 shows the employee page, there is two buttons: 1- collect donations where the employee can see a list of the donors with their information as shown in Figure 6. In addition, Button 2- Distribute donations where the employee can see a list of the needy people with their information. In the same page.



Maher
M7sn

Info

Name: Maher
Date: May 4, 2016
Time: 01:16:00
Type: Food

Fig. 6. Android app donor list with their information

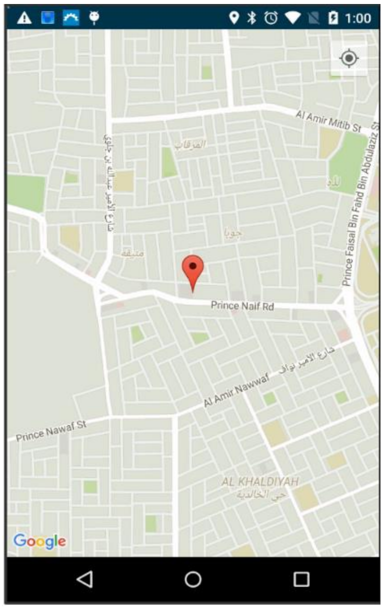


Fig. 7. Android app map page

Figure 7 show the map page where the employee can see the donor's location/needy people's location. After the employee click Collect/Distribute button he will see the donor/needy person information as shown in the previous Figure, and when the employee clicks on Drive button the map page will open where the employee can easily see and drive to the donor/needy person. In the same time, the map page used to see the donation box location in order to collect the donations.

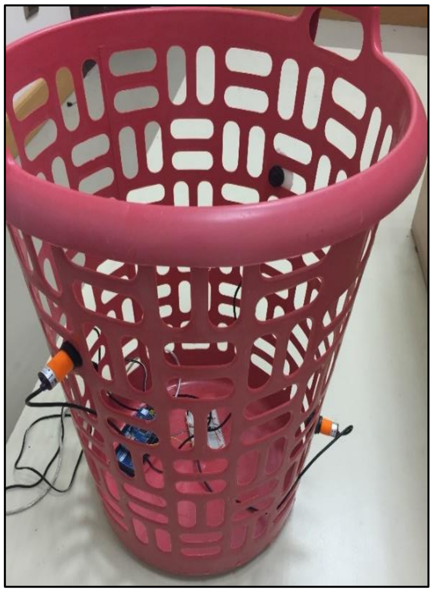


Fig. 8. Smart donation box prototype

Figure 8 shows the prototype of the Smart Donation Box (SDB) which all the component of the system is connected.

VI. Conclusion

The main goal of the Smart Charity (SC) system is to help the society in a modernized way using new era technologies to make donations and charity operations more efficient. SC introduced smart solution for the charity organizations and donation boxes, mostly placed around the city in the Kingdom of Saudi Arabia to turn them into Smart Donation Boxes (SDB) by using IoT, which will enable the existing donation boxes to notify to the charity organizations, when the collected donations reached to certain levels. In addition, SC helps the donors to donate with ease and comfortable way. SC also provides help to the needy people by forwarding them several notifications linked to the charity without feeling need to visit physically to the charity organizations. This system further can be extended to multiple smart gadgets platforms such as IOS, Windows etc. and can be extended to the rest of the world to enhance the ease and efficiency of charity organizations using IoT.

Acknowledgment

First of all, we would like to take this chance to express our significant appreciation to College of Computer Science & Information Technology (CCSIT) King Faisal University (KFU) for the support and environment provided to complete our system and implement it as a pilot project. We also extend our gratitude to Dr. Noor Zaman for his supervision, exemplary guidance, monitoring and constant support.

References

[1] Alber charity organization & Dar_Alkhair charity organization. Access at 1-Sep-15 Retrieved from http://www.albr.org/ and http://ahsaber.org/?page_id=126104 page: 55

[2] Coetzee, Louis, and Johan Eksteen. "The Internet of Things-promise for the future? An introduction." In IST-Africa Conference Proceedings, 2011, pp. 1-9. IEEE, 2011.

[3] Saudi food bank (2015). Access at 9-Sep-15 Retrieved from http://saudifoodbank.com

[4] Makkah charity (2015). Access at 14-Sep-15 Retrieved from http://www.mc.org.sa/

[5] NemahKeep (2013). Access at 26- Sep -15 Retrieved from http://www.hifz.org/

[6] Muhammad Sajidur Rahman, Khondoker Asif Akter, Shakil Hossain,Anjon Basak, Syed Ishtiaque Ahmed. "Smart Blood Query: A Novel Mobile Phone Based Privacy-aware Blood Donor Recruitment and Management System for Developing Regions".

[7] Fatimah AlDossari, Manal AlMubarak ,Marwa AlBukhowa and Maryam AlSaif, Noor Zaman . “Life Donors: saving lives by using current era smart technologies” Journal of Information & Communication Technology Vol. 9, No. 2, (2015) 55-76

[8] Abdur Rashid Khan and Muhammad Shuaib Qureshi. "Web-Based Information System for Blood Donation". International Journal of Digital Content Technology and its Applications Vol. 3, No.2 (2009).

[9] Sara A. Hashim, Afnan M. Al-Madani, Shatha M. Al-Amri, Abeer M. Al-Ghamdi, Bayan S. Bashamakh. NahlaAljojo, PhD. "Online Blood Donation Reservation And Management system In Jeddah". Life Science Journal Vol 11, No.8, (2014)

Virtual Observation System for Earth System Model: An Application to ACME Land Model Simulations

Dali Wang, Fengming Yuan
Climate Change Science Institute
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA

Yu Pei, Cindy Yao
Department of Electric Engineering
and Computer Science
University of Tennessee, Knoxville, TN 37996, USA

Benjamin Hernandez
National Center for Computational Science
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA

Chad Steed
Computational Science Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA

***Abstract*—Investigating and evaluating physical-chemical-biological processes within an Earth system model (EMS) can be very challenging due to the complexity of both model design and software implementation. A virtual observation system (VOS) is presented to enable interactive observation of these processes during system simulation. Based on advance computing technologies, such as compiler-based software analysis, automatic code instrumentation, and high-performance data transport, the VOS provides run-time observation capability, in-situ data analytics for Earth system model simulation, model behavior adjustment opportunities through simulation steering. A VOS for a terrestrial land model simulation within the Accelerated Climate Modeling for Energy model is also presented to demonstrate the implementation details and system innovations.**

Keywords—Earth System Modeling; Accelerated Climate Modeling for Energy; In-Situ Data Analytics; Virtual Observation System; Functional Unit Testing

I. Introduction

Over the past several decades, several Earth system models (ESMs) have been developed to understand Earth system dynamics and to project future climate scenarios. Among these ESMs, the Accelerated Climate Modeling for Energy (ACME) model, funded by the US Department of Energy (DOE), is a national effort to address the challenging and demanding climate-change research imperatives. Due to the complexity of EMSs in both model design and software implementation, the validation and verification of the Earth system process with EMSs are quite challenging, especially at the scales and levels of organization wherein many relevant field measurements and experiments are made (Wang et. al., 2014a). Scientists routinely use post-simulation approaches to analyze results. These include visual exploration to detect anomalies or interesting patterns and statistical data analysis for further investigation. Generating data for post-simulation earth system process investigation quickly becomes a cumbersome task once a simulation reaches a fairly large scale with a huge amount of data and daunting input/output cost. For these reasons, an interactive, run-time simulation monitoring system, or a virtual observation system (VOS), is needed. In this paper, author first scribe key functions of a VOS and then describe its major components based on advanced computing technologies (such as compiler-based software analysis, automatic code instrumentation, and high-performance data transport). At last, for the demonstration purpose, authors present implementation details on a VOS for a terrestrial land model that is the ACME Land Model (ALM) which is a process-based model with a collection of key bio geophysical and biogeochemical functions that represent the energy-water-biogeochemical interactions between the atmosphere and the terrestrial landscape. The VOS software system for ALM provides the capabilities of real-time observation and in-situ data analytics for model simulation.

II. Virtual Observation System Design

Key functions of A VOS are 1) to setup a “watch point” for a specific physical-chemical-biological function and 2) to capture the input and output data streams of a target function. Therefore, users can quantify the relationship between input and output data streams of a target function and identify variables that are can be observed at desired sampling frequencies. This information can be used to guide data collections in real world observation systems. A VOS also provides interactive tracking capability over user-selected key model variables throughout model simulation, so that users can “observe” changes in model variable values, and explore the relationship among Earth system functions (related to these user-selected model variables) over a specific spatial-temporal domain.

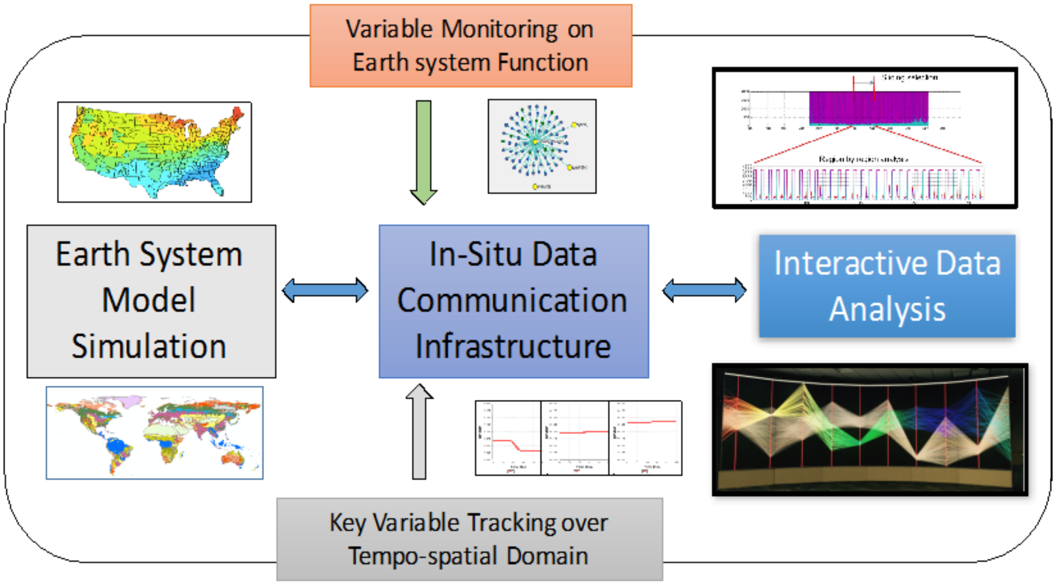

Variable Monitoring on Earth system Function
Earth System Model Simulation
In-Situ Data Communication Infrastructure
Interactive Data Analysis
Key Variable Tracking over Tempo-spatial Domain

Fig. 1. Major software components of a VOS, including software analysis and code instrumentation, in-situ data infrastructure, and interactive data analysis. Two typical uses of the VOS are function-specific data monitoring and variable tracking throughout the simulation

Figure 1 shows two typical uses of a VOS. First, the VOS allows users to define a specific function (an individual subroutine or a group of related subroutines) and an observation period, then the VOS collects input and output data streams of the target function and transports these data out of the simulation system for visualization and analysis. Second, the VOS helps to track specific key model variables throughout the simulation system over a user-defined period. Figure 1 also illustrates the major components of a VOS, including software analysis and code instrumentation, in-situ data communication infrastructure and interactive data analysis.

A. *Software analysis and automated instrumentation*

The main purpose of this VOS component is to collect information on software structures and workflow. Authors adopted a similar workflow procedure used in a scientific function test platform (Wang et. al., 2015, 2014b; Yao et. al., 2016). The procedure has several steps: First, authors use software dependency analysis to identify methods to reduce software dependency on parallel computing and external libraries. This step simplifies the model software dependency by using production compilers without an optimization option. Next, authors perform a compiler-assisted workflow analysis to capture the internal data structure and scientific workflow of the simulation source code. For a given function or module, authors use a programming language parser to analyzes the source code, break it into tokens, and store the program internally as an abstract syntax tree (AST). Then, authors conduct recursive name resolution through the AST to capture the input and output data streams of a target function in the simulation source code. Finally, authors instrument code segments into the source code to pack all the data of interest into a continuous memory buffer ready for in-situ data infrastructure. Since the majority of EMSs are developed in Fortran, authors are working on the integration of a kernel extraction tool (Kim, et. al., 2016), which is built on top of a Python Fortran parser, for automatic code instrumentation. The process is shown in Figure 2.

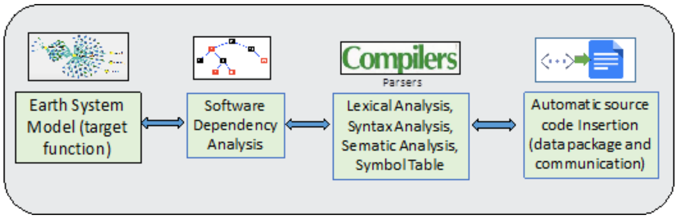

Earth System Model (target function)
Software Dependency Analysis
Compilers
Parsers
Lexical Analysis, Syntax Analysis, Sematic Analysis, Symbol Table
Automatic source code Insertion (data package and communication)

Fig. 2. General procedure for software analysis and automated instrumentation within a VOS

B. *In-situ data communication infrastructure*

The main function of this VOS component is to provide high-performance data communication capability for transferring the data of interest out of simulation system for external analysis. The data infrastructure allows users to inspect variable values in real time during model simulation. In the current effort, VOS in-situ data infrastructure is built on the Common Communication Interface (CCI) (Atchley et al, 2011). The CCI project is an open-source communication interface that aims to provide a simple and portable Application Programming Interface (API), high performance and scalability for the largest deployments, and robustness in the presence of faults. The in-situ data infrastructure consists of three segments: data generation, data staging, and data analysis (Figure 3). In the VOS, the data analysis segment first creates CCI channels to which the data generation segment (instrumented simulation code) can connect. Once the connection is established, users can then pass simulation parameters (function and variable names, time interval, and location, etc.) to instrumented simulation code. Once the simulation runs to the user-defined time interval, the instrumented simulation code packs all the relevant data into a buffer and uses CCI's Remote Memory Access (RMA) methods to send the data over the network to the data analysis segment. The data analysis segment always listens on its own CCI channel. When the data arrives, the analysis segment unpacks the data for follow-up data processing and analysis.

Considering that large data volume needs to be transferred into data analysis, VOS data infrastructure also includes a data staging area that allows data caching for input/output operations and low-latency data queries. The data staging area also allows users to define functions and observation periods and track key model variables over simulation period. The main purposes of data staging are: 1) reduce potential data overload in the analysis side during model simulations and 2) then enable user-based queries and maintain interactive rates. Currently, the staging area is co-located with data analysis and visualization, and acts as a temporal storage area for data processing operations (e.g., storing, loading, extraction, transformation, or querying). Figure 3 shows the VOS in-situ data infrastructure with a staging area.

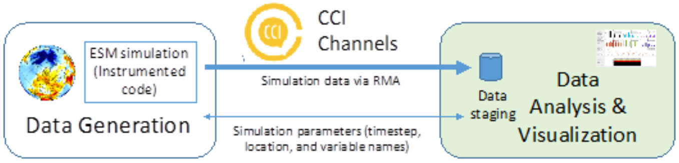

ESM simulation (Instrumented code)
Data Generation
CCI Channels
Simulation data via RMA
Simulation parameters (timestep, location, and variable names)
Data staging
Data Analysis & Visualization

Fig. 3. In-situ data infrastructure with a data staging area inside the data analysis component

C. *Interactive data analysis*

This VOS component provides a front end with which users can perform three main tasks: 1) choose the ecosystem functions and time interval for monitoring, 2) interactively visualize the results of predefined “watch” points throughout simulation, and 3) steer the simulation accordingly, if necessary. The data analysis component also directly communicates with the staging area to conduct query submission and data retrieval based on the user interactions.

From the technical perspective, this component contains three modules: 1) a graphic user interface (GUI) that allows users to perform these three main tasks, 2) an interactive data visualization engine that plots physical-chemical-biological interactions produced by the simulation, and 3) a communication interface with a staging area which in turn connects to the instrumented simulation code.

In the study, the GUI is built using Qt and the data visualization engine is developed using the Visualization Toolkit (VTK), which utilizes the underlying graphical processing unit (GPU) for faster rendering. Multicore CPU processors are used to handle data transfer. After receiving the buffer from CCI, the engine converts the data into vtkTable data structure for visualization. The buffering mechanism based on data staging allows users to select time steps for visualization. The visualization engine employs a client-server model, so that while the VTK server is located alongside the simulation for faster data transfer, the actual client display windows can be on any remote machine. This feature greatly increases the portability and usability of the system.

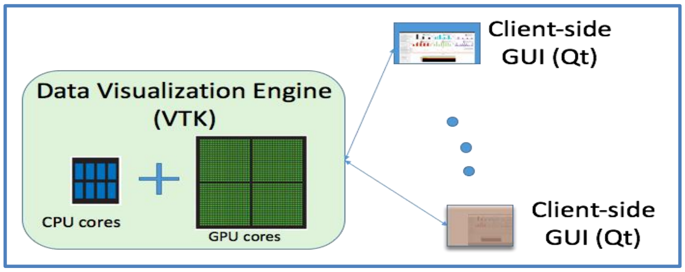


Fig. 4. Key components of VOS data visualization, which utilizes hybrid hardware and provides cross-platform GUIs

III. VOS FOR ALM: CASE DEMONSTRATION

In this section, authors demonstrate a VOS for the ALM simulation over the Next Generation Ecosystem Experiments Arctic site (NGEE-Arctic, http://ngee-arctic.ornl.gov), located at the Barrow Ecosystem Observatory (BEO) in Barrow, Alaska. In this experiment, ALM was configured as a point-mode offline simulation to investigate terrestrial ecosystem responses to specific atmospheric forcing over a single landscape grid cell at Barrow (Yuan, et. al., in preparation). For the demonstration purposes, the observation system is used to track all the variables in and out of a CNAllocation module within ALM. The CNAllocation function is developed to allocate key chemical elements (such as carbon, nitrogen and phosphorus) of a plant in a terrestrial ecosystem.

The software architecture diagram of the VOS for ALM using the CNAllocation module is illustrated in Figure 5.

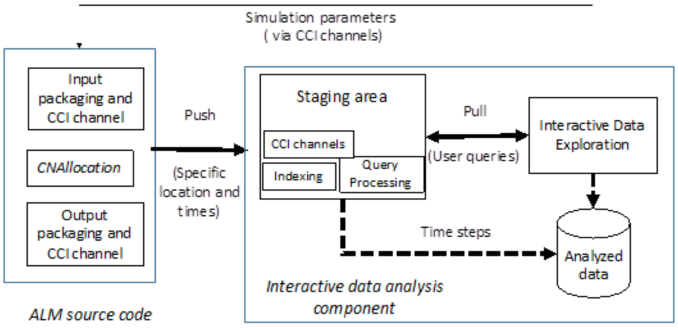
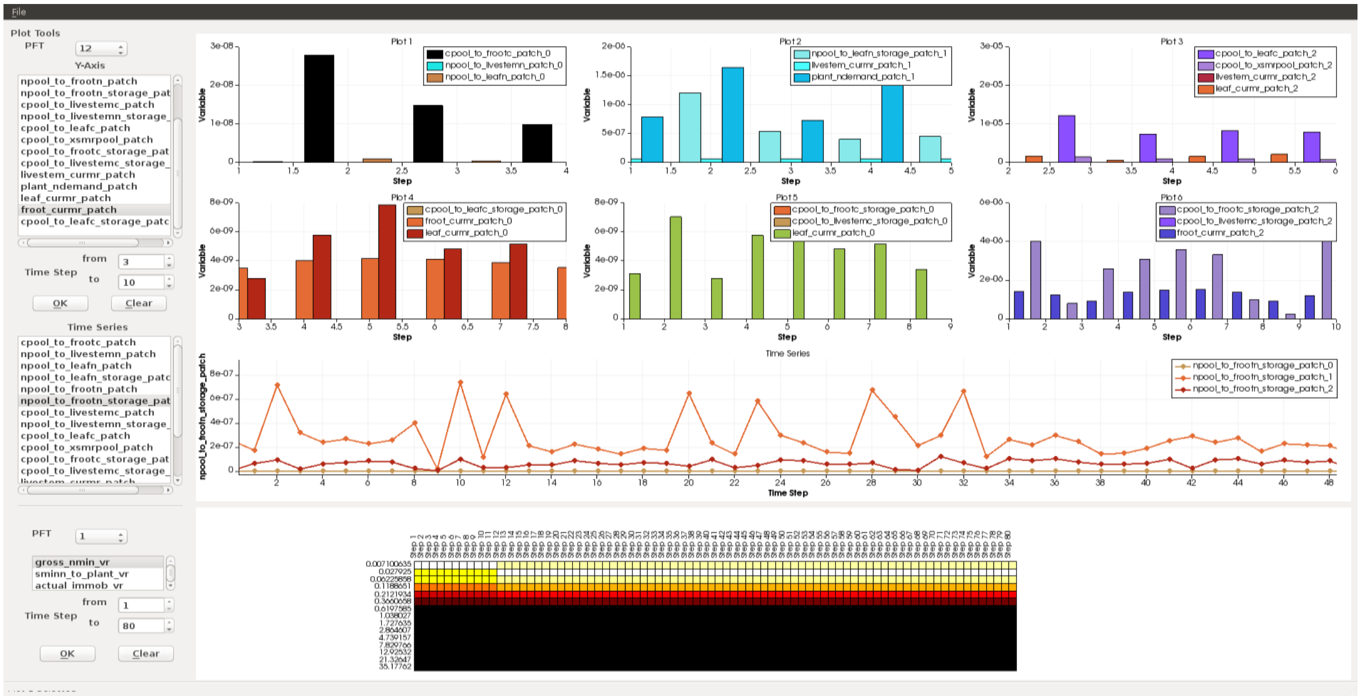

Simulation parameters (via CCI channels)
Input packaging and CCI channel
CNAllocation
Output packaging and CCI channel
ALM source code
Push
(Specific location and times)
Staging area
CCI channels
Indexing
Query Processing
Pull
(User queries)
Interactive Data Exploration
Time steps
Analyzed data
Interactive data analysis component

Fig. 5. The schematic software architecture diagram of the VOS for the ACME Land Model

As shown in Figure 5, code segments are instrumented into the source code to capture and pack the input and output data streams of the targeted module, CNAllocation. The code segments also contain functions that invoke the in-situ data communication infrastructure, including CCI channel and data buffer. The VOS has a staging area that also contains a CCI channel and data buffer. The staging area is accessible from a data exploration subcomponent. Authors first start the interactive data analysis component, which takes user-specified parameters (such as time interval, or a subset of variables) and then listens to the CCI connection requests from the simulation side. Next, authors start the instrumented ALM simulation code. When the simulation code runs to the user-defined time steps, the instrumented code packages all the relevant data into a buffer and then sends the buffer to the interactive data analysis component over the network. The data analysis component always listens on its CCI channel. When data arrive, the data analysis component unpacks the data in the staging area for follow-up data processing and analysis.

The GUI for CNAllocation data analysis and exemplar simulation data streams is illustrated in Figure 6. The first two rows show different bar plots of carbon and nitrogen allocation variables for a plant type over a specific range of time steps. The third row displays a time series from given carbon and nitrogen allocation variables; this graph allows users to track the behavior of target variables during the simulation. The fourth row includes a heat map for plotting variables having a 2D domain. Finally, the left panel shows the complete variables and time step selection.



Plot Tools

PFT 12

Y-Axis

npool_to_frootn_patch
npool_to_frootn_storage_pat
cpool_to_livestemc_patch
npool_to_livestemn_storage_
cpool_to_leafc_patch
cpool_to_xsmrpool_patch
cpool_to_frootc_storage_pat
cpool_to_livestemc_storage_
livestem_curmr_patch
plant_ndemand_patch
leaf_curmr_patch
froot_curmr_patch
cpool_to_leafc_storage_patc

Time Step from 3 to 10

Time Series

cpool_to_frootc_patch
npool_to_livestemn_patch
npool_to_leafn_patch
npool_to_leafn_storage_patc
npool_to_frootn_patch
npool_to_frootn_storage_pat
cpool_to_livestemc_patch
npool_to_livestemn_storage_
cpool_to_leafc_patch
cpool_to_xsmrpool_patch
cpool_to_frootc_storage_pat
cpool_to_livestemc_storage_

PFT 1

gross_nmin_vr
sminn_to_plant_vr
actual_immob_vr

Time Step from 1 to 80

File
OK
Clear
OK
Clear
Plot 1
cpool_to_frootc_patch_0
npool_to_livestemn_patch_0
npool_to_leafn_patch_0
Plot 2
npool_to_leafn_storage_patch_1
livestem_curmr_patch_1
plant_ndemand_patch_1
Plot 3
cpool_to_leafc_patch_2
cpool_to_xsmrpool_patch_2
livestem_curmr_patch_2
leaf_curmr_patch_2
Plot 4
cpool_to_leafc_storage_patch_0
froot_curmr_patch_0
leaf_curmr_patch_0
Plot 5
cpool_to_frootc_storage_patch_0
cpool_to_livestemc_storage_patch_0
leaf_curmr_patch_0
Plot 6
cpool_to_frootc_storage_patch_2
cpool_to_livestemc_storage_patch_2
froot_curmr_patch_2
Variable
Step
Time Series
npool_to_frootn_storage_patch
npool_to_frootn_storage_patch_0
npool_to_frootn_storage_patch_1
npool_to_frootn_storage_patch_2
Time Step

Fig. 6. GUI of the VOS data analysis of the CNAllocation functions within the ACME Land model. Users can zoom in or out to inspect different time steps or drag on any plot to highlight certain variables

IV. Conclusion

Authors have demonstrated an approach to develop a virtual observation system (VOS) for Earth system models. Authors also have implemented a VOS for the ACME Land Model using a single point-mode simulation case. By taking advantage of compiler-based software system analysis, automatic code instrumentation, and high-performance in-situ data transport, the VOS provides unique capabilities to investigate Earth system behaviors in a unique way. The VOS is designed based on non-intrusive observation principles; it preserves all the original software data flow and function calls. The VOS also allows scientists to interactively select targets of interest, such as key variables, functions, or specific break points for a simulation. Modelers can focus on investigating model behaviors without dealing with complex code instrumentation and large data handling on high-performance computing platforms. Future work will focus on two directions: 1) extending two-way communication mechanism to improve the efficiency of data collection and 2) integrating with external big data visual analysis toolkits (such as EDEN (Steed et al., 2013)) and existing advanced statistical analysis packages (such as R (Horsburgh et al., 2014) and Matlab (Pianosi et al., 2012)). The latter requires further development of data staging nodes within the system. In this extension, Dataspaces library (Docan et al., 2012) could be used to allocate and manage data staging nodes and handle push and pull operations between the VOS components, whereas Fastbit library (Wu, 2005) could be used for data indexing and query processing within these nodes, and CCI can still provide a two-way communication between simulation and analysis components to enable simulation steering.

V. Software Availability

VOS has been tested on a variety of computing environments (from desktop to high-performance computer cluster). VOS uses the software parsing and instrumentation capability developed through a functional unit testing platform for ALM (in Fortran). The functional testing platform uses compiler-based technology for software analysis and code instrumentation. The source code of the functional unit testing platform is located at a unit testing repository within bitbucket. (https://bitbucket.org/cindy387/clm85/src/cfa8d8faa43a21dcdde9b8750a9816a92477a361/?at=DEMO). Currently, the in-situ data infrastructure code is developed based on CCI libraries (in C), and is located at a CCI-in-situ repository in bitbucket.

(https://bitbucket.org/cindy387/clm85/src/83f7ade49968afef18 dd944560a343adbd6a3810/?at=In-situ). The visualization package can be found at https://bitbucket.org/benjha/dataviz-acme-land-model.

ACKNOWLEDGEMENT

This research was funded by the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research (BER) program and Advanced Scientific Computing Research (ASCR) program, and by an ORNL LDRD project (#7409). This research used resources of the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at Oak Ridge National Laboratory (ORNL), which is managed by UT-Battelle LLC for the Department of Energy under contract DE-AC05-00OR22725.

REFERENCE

[1] Atchley, S., Dillow, D., Shipman, G., Geoffray, P., Squyres, J., Bosilca G., and Minnich, R., 2011, The common communication interface (CCI) in the 19th IEEE Symposium on High Performance Interconnects (HOTI), Santa Clara, CA, August 23-25, 2011.

[2] Docan, C., Parashar, M., and Klasky, S., 2012. “DataSpaces: an interaction and coordination framework for coupled simulation workflows”. Cluster Computing 15(2): 163-181, doi: 10.1007/s10586-011-0162-y

[3] Horsburgh, J.S., Reeder, S. L., Data visualization and analysis within a Hydrologic Information System: Integrating with the R statistical computing environment, Environmental Modelling & Software, Volume 52, February 2014, Pages 51-61, ISSN 1364-8152, http://dx.doi.org/10.1016/j.envsoft.2013.10.016.

[4] Pianosi, F., Sarrazin, F., Wagener, T., A Matlab toolbox for Global Sensitivity Analysis, Environmental Modelling & Software, Volume 70, August 2015, Pages 80-85, ISSN 1364-8152, http://dx.doi.org/10.1016/j.envsoft.2015.04.009.

[5] Steed, C. A., Ricciuto, D.M., Shipman, G., Smith, B., Thornton, P.E., Wang, D., Shi, X., Williams, D. N., 2013, Big data visual analytics for exploratory earth system simulation analysis, Computers & Geosciences, Volume 61, December 2013, Pages 71-82, ISSN 0098-3004, http://dx.doi.org/10.1016/j.cageo.2013.07.025.

[6] Wang, D., Schuchart, J., Janjusic, T., Winkler F., and Xu, Y.,2014a. Toward better understanding of the Community Land Model within the Earth System Modeling Framework, in: Abramson, D; Lees, M; Krzhizhanovskaya, W., Dongarra, J; Sloot, P.M.A. (Eds.), Procedia Computer Science, 14th Annual International Conference on Computational Science, Cairns, Australia, 2014, Procedia of Computer Science, Volume 29, 2014, Pages 1515–1524, 10.1016/j.procs.2014.05.1375.

[7] Wang, D. Xu, Y., Thornton, P., King, A., Gu, L., Steed, C., Schuchart, J., 2014b, A functional testing platform for the Community Land Model, Environmental Modeling and Software, 2014, Volume 55, Pages 25-31, 10.1016/j.envsoft.2014.01.015

[8] Wang. D., Wu, W., Janjusic, T., Xu, Y., Iversen, C., Thornton, P., Krassovski, M., 2015. Scientific functional testing platform for environmental models: An application to the Community Land Model, International Workshop on Software Engineering for High Performance Computing in Science, 37th International Conference on Software Engineering, May 16-24, 2015, Florence, Italy. Doi: 10.1109/SE4HPCS.2015.10

[9] Wu, K., 2005. “FastBit: an efficient indexing technology for accelerating data-intensive science” J. Phys.: Conf. Ser. 16 556-560, doi: 10.1088/1742-6596/16/1/077

[10] Yao, Z., Jia, Y., Wang, D., Steed, C., Atchley, S., 2016, In situ data infrastructure for scientific unit testing platform1, in: Connelly, M. (Ed.), Procedia Computer Science, Volume 80, 2016, Pages 587-598, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2016.05.344.

[11] Youngsung Kim, Y., Dennis, J., Kerr, C., Kumar, R., Simha, A., Baker, A., Mickelson,S., 2016, KGEN: A Python tool for automated Fortran kernel generation and verification, in: Connelly, M. (Ed.), Procedia Computer Science, Volume 80, 2016, Pages 1450-1460, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2016.05.466.

[12] Yuan, F., Thornton, P. E., Xu, ,X., Sloan, VL., Iversen, C., Rogers, A., Yang B., and Wullschleger S. D., (2016). Modeling analysis of assimilate partitioning between storage and pools of multiple plant function types for simulating carbon cycles in Arctic coastal tundra ecosystem at Barrow, Alaska. JGR-biogeoscience (in preparation)

RTS/CTS Framework Paradigm and WLAN Qos Provisioning Methods

Mohamed Nj.
Dept. Computer Communications & Network Systems
FTMK – Teknikal Universiti Malaysia
Melaka, Malaysia

N. Suryana
Dept. Software Engineering
FTMK – Teknikal Universiti Malaysia
Melaka, Malaysia

S. Sahib
Dept. Computer Communications & Network Systems
FTMK – Teknikal Universiti Malaysia
Melaka, Malaysia

B. Hussin
Dept. Industrial Computing
FTMK – Teknikal Universiti Malaysia
Melaka, Malaysia

***Abstract*—Wireless local area network (WLAN) communications performance design and management have evolved a lot to be where they are today. They went through some technology's amendments and innovations. But, some performance tools remained almost unchanged and play a fundamental role in contemporary networking solutions despite the latest innovations higher influence on their indisputable and important function. That is the case with Request to send (RTS) and consent to receive (CTS) protocols. They are among the former technologies, which helped for transmission control with better performance in WLAN environment. They are so important, particularly since the advent of sensitive data networking (e.g. internet telephony, audio and video materials distribution) over the internet protocol (IP). Up to recent years following today's multimedia WLAN based networks deployment trends, RTS/CTS) contributed to provide networks with some expected good performance levels prior to the discovery of more sophisticated methods for this purpose (i.e. performance enhancements). And yet, one may question whether the new technologies have rendered RTS/CTS frameworks obsolete; or are they now used only for some specific network applications traffic management? This articles review attempts to comprehensibly study some of the research works, which have had interest in RTS/CTS mechanism as tools for WLAN applications performance support. Various researches have studied these tools from their early innovation as network node's built-in component, through different frameworks associated with WLAN legacy (IEEE 802.11) MAC protocols. This paper analyzed RTS/CTS initial implementation as mere network performance solution from packets' collision avoidance perspective; and then for transmission delay due to hidden nodes and their false deployment. The article closes up on a critical analysis on the possible long time contribution of these protocols into integrated schemes based WLAN QoS performance design.**

Keywords*—RTS/CTS; MAC; Internet; Telephony; video; real-time; loss; multimedia; WLAN; mechanism; performance; protocols, collision; framework; transmission; reception; flow control; handshake; MANET; BSS; IBSS; QoS

I. INTRODUCTION

Request to send and consent to receive (RTS/CTS) are one of the main elements of flow control on network communicating nodes, which acts at such individual device level as a special gateway for data transmission (Sending and reception). They are sockets/ports embedded into almost every network's end-nodes. Their mechanism's function is very important and thus more valuable on the client side performance's management. In fact, many research works have been done about local area network (LAN) and wireless LAN (WLAN) management using RTS/CTS as performance support to their different service applications. WLAN is an ever great platform tool for wired networks (LAN) extension to wireless and then mobile networks of all kinds. RTS/CTS usefulness is reported in various studies. First, they are a feature of the IEEE 802.11 WLAN technology having initially for main function to control station access to the shared medium. A correct implementation (Timing on/off and threshold setting) of RTS/CTS lets user adjust/regulate the WLAN packets transmission relatively to the operation environment [1]. Practically, these sockets on the client's device enable timing packets transmission after making first a request (RTS frame sent) and receiving reply (CTS frame received) from a peer network node [2, 3]. They are used as an optional technique [4] in wireless LAN legacy (IEEE 802.11) for transmission control between clients and the access-point (AP) [2, 4]; they are so known also as best tools in negotiating or ensuring bandwidth [2, 4, 6] prior for a client transmitting its data.

In WLAN environment, radio interference can occur relatively to the terminals' location and position to the AP; this is generally known as hidden terminal issue [**3, 4**]. Further details to this issue are under sections three and four of this article. A systematic consequence of this situation is the packets collision at the AP for attempted transmissions between any client and the hidden one. However, collision is proven in literatures as the basic source for data loss. In turn, this will lead to end-to-end (E2E) throughputs decrease as negative effects with undesirable delays (e.g. in phone calls). To ensure WLAN good performance, IEEE 802.11 media access control (MAC) uses either of the following two techniques against the interference occurrence: RTS/CTS mechanism and the physical carrier sensing media access

(CSMA). RTS/CTS handshake is a virtual carrier sensing known as perfectly able to reduce interference and related consequences [2, 3, 4, 5]. RTS/CTS thresholds proper settings along with a good adjustment of wireless local area network (WLAN)'s clients within the access-point (AP) transmission range are among the strategic methods to obtain good control for great performance of the network [2, 3, 4**]**.

Comparing the two above mentioned techniques, the second is proven less effective in solving for the interference issues in WLAN. In fact, this remark is made relatively to the RTS/CTS potential capability for the same task of controlling clients transmission to avoid collision occurrence [6]. In other words, RTS/CTS efficiency is subject to the use with each of WLAN operation modes (i.e. Basic set of services (BSS) and Independent BSS (IBSS)). Despite the wireless technologies development for different application's perspectives, yet the above stated matter still holds much believe for RTS/CTS support capability. For instance, in sensor wireless network environment (IBSS-WLAN) the use of geographic positioning system (GPS) as part of integrated solutions support sounds only very advanced hints, but limited in efficiency. Overall, attempted solutions can be tried and only manually [7]. In general, WLAN contemporary deployment method allows users exploring flexibly the networks, including telecommunications network operators to rate by the wireless users [8, 9]. However, the typical ways of deploying WLAN for a same purpose or more include (a) LAN extension, (b) cross-building interconnect, (c) nomadic access, and (d) ad-hoc networking [10]. Despite their performances solution differences, they experience hidden stations or nodes localization and position issue. The deployment objectives and advantages out of various studies' review show almost the same findings [11]. Namely as LAN extender and enabler for users mobility when connected; currently the best of internet gateways (Portal) whether indoor or outdoor use due to many related benefits: cheap cost and easy deployment for requiring only little IT-knowledge for systems configuration; moves with and access to various applications and services regardless of server location and time; etc. Wireless operations made it possible short and long distance communications, including the unrealizable projects in wired networking [12]. Hence, such a great evolution in communications technology has allowed a vast majority of little and even zero-level computer literacy people to get exposed to both the advance telecommunication system (e.g. Internet, IP based telephony and smart-telephony) and wireless mobile networking (e.g. mobile WLAN, smartphone uses). This group of consumers in fact makes up a considerable number of consumers in internet and telecoms market especially in third world countries. Hence, this situation calls upon the industry and service providers (NSP) attention in providing them with more easier means for network troubleshooting during their access to network.

The foremost and particular learning from above sources and related studies was/is about the importance of wireless networks performance management using lower level tools to support possible automated solutions from all other network's levels. This article attempts to highlight the lack of enough discussions on such provisions in QoS or performance based new research papers. This remark includes a limited access offer to those tools' features, with RTS/CTS threshold settings as special case of concern [33, 34]. Some recent QoS surveys based articles have helped much verify this remark, as analysed here in sections three and four.

II. WLAN Operation and Performance Issues

Request to send and consent to receive (RTS/CTS) mechanism is among the tools in early network technology embedded on networking communications hardware. WLAN standard IEEE 802.11 contains RTS/CTS protocols to control clients access to shared-medium according to configured threshold [5, 6]. In fact, WLAN for open access/hotspot (e.g. Cafés, Offices, Hostel, etc.) faces multiple users at various locations/positions. Thus, in such complex situations, traditional equipment is merely inadequate to fairly deal with the rate of clients' service demands and generated interferences. However, some friendly management tools are available but only on expensive products. They provide users with RTS threshold settings in WLAN radio network interface cards (NIC) and AP interfaces [3, 6, 5].

A. WLAN Basic Operation Modes Overview

Wireless LAN is obviously considered as the base network for modern mobile networks access in home, office and medium organization's services.

1) WLAN Types of Deployment

The legacy (IEEE 802.11) provides two operation modes. There is infrastructure mode*:* - in which wireless stations remain in mobile communications but depend strictly on the AP radio coverage range, which act as bridge to other subnets, LAN and the internet. And the other is ad hoc mode: - for which mobile wireless stations interact directly in a peer-to-peer manner. In both operation modes, WLAN MAC protocols configurations for physical layers and medium access are the same, despite some modifications in 'mobile ad-hoc network' (MANET) case [13, 14]. However, MANET systems experience different topologies relatively to the mobile nodes location and positions to each other over time. Thus, such a change situation is a cause to their operation problems and solutions difference [13]. Hence, various research studies showed that RTS/CTS mechanism is not considered suitable for MANET environment but instead CSMA/CA based distributed MAC in each terminal [5, 6, 15].

2) Mobile Ad-hoc Network

Mobile ad-hoc networks (MANET) known also as independent basic set of services (IBSS) based WLANs are the model suitable in the regions impracticable for ordinary wired network deployment. In MANET, mobile stations are self- networks deployment [14]. In MANET, mobile stations are self-organized [13, 14] and therefore they need a distributed MAC. In fact, shared MAC or shared bandwidth mechanism is more convenient in infrastructure based WLANs [13]. The solution scheme for hidden stations/collisions in MANET preferably [13] use the called 'carrier sense multiple access with collision avoidance' (CSMA/CA). As a lesson from above references, RTS/CTS successful use requires WLAN manager's attention for proper additional control and necessary adjustments. Hence, since

MANET system does not possess such centralized management, therefore CSMA/CA and variant mechanisms make senses as choice for applicable solutions scheme in such type of wireless networks.

3) *WLAN'S Clients Communications Issues*

Whether in infrastructure based or in MANET operation-mode, WLAN faces either of the following performance challenges: (a) bandwidth limitation, (b) radio interferences, (c) transmissions collision, (d) congestion, and (e) outcome problem. Each of these problems has a critical impact on the WLAN applications performance, especially on the VoIP applications (e.g. Calls, videos, and other real-time transactions application). In general, interferences can cause packets collisions due to hidden receiver-station (Fig. 1) leading to E2E delivery decrease. In the other hand collisions induce congestion, which cause throughputs reduction and more in delay. The increase in delivery delay originates from retransmissions and its execution time [3, 4, 6]. In overall, the network operation ends up with a poor services quality resulting from performance degradation due to the above stated facts [6, 15].

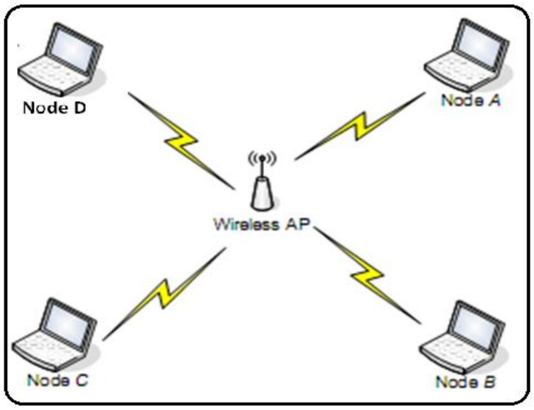


Fig. 1. WLAN hidden 'receiving' terminal issue illustration

However, in wireless network, hidden receiving stations problem is among the first severe causes of WN performance degradations [13] and thus low QoS [3, 4]. Technically wireless network's invisible and exposed nodes problem is generally due to either radio wave interferences (e.g. Case with huge of users in hotspots and condominiums), or short range between contending nodes [6, 13, 15]. Therefore, network topology must be accounted at least in maintenance process; that is because of the position nodes location involvement into hidden nodes problem. In fact, the use of RTS/CTS mechanism helps fairly tackle the unheard or hidden station's issue. So doing, this strategic solution leads in turn to another similar problem known as exposed nodes problems [4, 14]. That is, once any of the exposed nodes (e.g. Fig.1 'A' node) hears/senses the CTS exchange originated from a station (e.g. Fig.1 node 'C') to which it wanted also to transmit packets to, this node ('A') will simply drop its own packets without genuine proof for probability of loss [14] in the case that those packets could have been sent. And, such false abstentions consequently will gradually reduce the network throughputs delivery [2-6, 15], then the performance and finally the QoS on top of all.

The technical causes to hidden receiving node are discussed under solution's design (section IV). However, the scenario portraying unheard terminal can be illustrated like in Figure 1. Any of the four stations can fail to connect with either of three others though they are logically interconnected through a same Wireless AP (WAP). There are almost three general cases with this problem as discussed in [15]: (a) – either all nodes cannot hear each other; (b)– or they are visible but in contention for resources each other; (c)– and else they are invisible stations and contention happens simultaneously.

As particular learning, "many past researches have proved that about 40% of packets loss in wireless networks occurred because of invisible terminals problem" [15]. Therefore, well-designed RTS/CTS threshold and its proper implementation remain the fundamental way to solve the issue and incumbent defects [1-6, 15, 35]. The analyses in recent studies among these sources are implicitly an alert for more consideration to this solutions framework due to some observed achievements. On the IEEE 802.11 physical carrier sensing, some alternative methods (e.g. Clear channel assessment (CCA), fragmentation, queuing discipline, etc.) to RTS/CTS have been explored and tested for comparison between their potential extents [36-38]. But, RTS/CTS still over performed and better promised more hopes for services applications requiring QoS [39, 40] .Thus, multimedia based network applications can expect more for great QoS performance supported by integrated solutions. This is understandable, since local network managers need to enable/disable RTS/CTS assignments where and when applicable. It is then a coordinated effort to maintain performance in addition to any QoS level obtained from the network service provider. Furthermore, such above highly rated and pertinent remark from many studies is significant enough to prove that RTS/CTS fundamentally are needed in support to any other single or multiple schemes based performance solutions.

B. *How Does RTS/CTS Mechanism Operate?*

RTS/CTS are another alternative WLAN MAC operations support [3, 6], which can be manually configure as a typical solution against wireless network (WN) frame exchange collisions [1-6, 15]. This mechanism comes into playing its function by enabling /disabling its thresholds based on the WN behaviour with respect to the throughputs decrease level as observed by the network managers [13, 14]. However, its inadequate settings can degenerate instead into the network performance failure if not implemented based on proper finding out of the WLAN behavior's survey and results analysis. The use of thresholds must then according to findings.

a) Application of RTS/CTS Mechanism

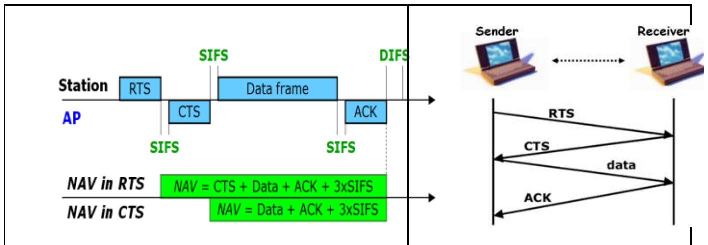
Functionally, in shared MAC medium, RTS/CTS mechanism enables controlling the WLAN client's frames exchange with others clients within same or outside subnets via the AP. The protocols handshake's algorithm (Table 1) uses one of the following control techniques – carrier sense multiple access with collisions detection (CSMA/CD), or carrier sense multiple access with collision avoidance (CSMA/CA) mechanism [5]. To enable or disable RTS/CTS mechanism means to 'activate/deactivate' its protocols system. In practice, this technically has to do with the action of ‘configuring the thresholds. These are parameter’s values that will decide and model the behavior of packets during their transmission process. Moreover, for many literatures, these control settings are performed on the wireless clients; for, these are either a source or an end node of the packets transmission [2, 6]. The thresholds are not applied on the AP [2, 3, 4]. Systematically, AP intervenes instead as a referee and dispatcher of frames between sources and destinations. While this control mechanism’s configuration interests more the client side [4, 5, 6], the AP by default learns from the clients operation. Then, to play its own role the AP quickly adapts to the clients traffic behavior, which depends on the applied RTS/CTS thresholds [2, 3].

RTS Enabled versus Disabled:

When a terminal’s RST/CTS are activated, it always holds down its packets from sending and will release them only after obtaining CTS frame from intended destination (Figure 2.2(a)). Thus, this process enables minimizing packets collisions occurrence and thus improving the network performance [2, 6]; see Figure 2 and Table 1.

RTS Disabled

When RTS is disabled on a terminal (e.g. AP, end-node/client) this latter relies on the WLAN MAC technique called physical carrier sensing (PCS) for that terminal’s packets transmission’s control [2, 3, 4, 6]. However, clear channel assessment (CCA) mechanism makes used of PCS threshold to check and decide for which among nodes contending for channel free can safely transmit. Thus many literatures [14, 36,41, 42] claimed for the efficiency of this method in fairly handling the hidden stations. In addition, almost all research studies on wireless ad-hoc networks favor the used of PCS versus RTS/CTS mechanism as solution to combat collisions occurrence.



[a] RST/CTS Scheme & Timeline

[b] RTS/CTS mechanism Four-Handshake Simplified

Fig. 2. RTS/CTS scheme in BSS based WLAN

TABLE I. BSS HANDSHAKE ALGORITHM (FIGURE 2.B)

1)	‘A’ ready to send frame to ‘B’: All stations’ waiting period/Initial →DIFS
2)	‘A’ ready & send request (RTS) to ‘B’ (via AP, controls and dispatching point)
3)	AP (Intercepts A’s request) issues CTS to all others (C, & D) with timeout
4)	‘B’ (responds via AP) sends frame ‘CTS’ to ‘A’
5)	‘A’ (Packets) send data to ‘B’ (via AP)
6)	‘B’ (ACK frame) send ACK to ‘A’ (via AP)

b) Protocols RTS/CTS Operation Explanations

Case for Shared MAC and Exclusive Single Access:

This refers to WLANs operating in BSS mode. The label “single exclusive access” [3, 6, 15] indicates the fact that one and only one station’s packets can be transmitted once the AP has declared the medium idle. That is clear since in this case the interactions between clients rely on the AP MAC ruling, contrary to MANET where everyone owns its MAC and BSS, despite of sharing open wireless network.

Handshake in BSS (Fig. 1&2):

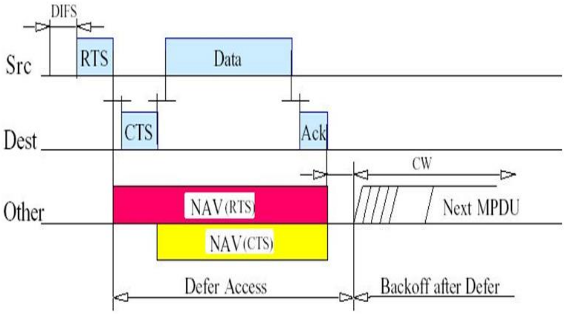
This scenario has 4 terminals (A, B, C &D). Say, “A” is ready to communicate with “B”. Thus, “A” first initiates it with a request (frame RST) via the AP. Then, AP reacts to this request with a CTS frame to all [2]; but, C & D will receive it along with timeout value (10-16µ seconds) [16] as “warning” about medium busy (i.e. they must hold-down their request if any for this period). Thus, “A” can release its frame toward “B” upon receiving the CTS. And then the session closes up with (“B”)’s acknowledgment (ACK) frame back to “A”.

For every BSS client, Figure (2.a) shows that RTS/CTS scheme relies mainly on both short inter-frame space (SIFS) and the network allocation Vector (NAV) for wireless medium access management [2, 3, 15, 17, 18]. SIFS controls the time interval (SIFS = 10 to 16 µs) between consecutive frames crossing the shared medium. NAV assures (up to 50µs, longer enough) [16] free medium use only for sender-receiver exchange of frames. Finally, the next frame (RTS) from one of two among stations initially put on queue will be the station having just for SIFS timeout; the one with DIFS timing will be set back on queue [17].

As learnt lesson, one can analyze these conditions associated with RTS issuing, receiving CTS before releasing packets and then ACK to packet’s sender ending a session. They are the proven facts on the efficiency of this mechanism against the collisions occurrence. Therefore, RTS/CTS handshaking offers enough control on the shared medium access [3, 6, 13, 15, 16, 17]. Various research studies warn about the unnecessary implementation of RTS/CTS. For, this can create the called “induce congestion”, resulting into some increase in overhead and thus a network performance dropping.

Case for Multiple Accesses MAC:

The reference [15] presented an example of handshake using multiple accesses with collision avoidance for Wireless (MACAW) technique. The packet transmission control in the MACAW is similar to the case discussed in Figure 2. But, MACAW applies instead a pattern of RTS/CTS/Data Sending (DATA/ACK / (DS)) for data transmission. A detailed presentation of these solutions is available in [15] study article.



RTS/CTS Handshake Random Access MAC protocols for MANET support

Fig. 3. RTS/CTS scheme in IBSS based WLAN or MANET

RTS/CTS handshake random access MAC protocols (Figure 3) are the scheme model suitably applicable on mobile stations in wireless ad-hoc network [16, 17, 18, 19]. Since every mobile station owns a MAC, the access to the medium is individually negotiated; thus random access, because of various attempts based on the back-off space and differ access behaviors. The references [14, 15, 16, 18, 19, 20, 21] discuss more in-depth about this scheme operations.

c) Managing RTS Induced and Normal Congestions

RTS induced congestion is linked to WLAN MAC layer operation. However, normal or systematic congestion happens on the TCP/IP based networks transmission as a result of buffers’ overflow [6]. Otherwise, a network overall congestion situations can be considered (Figure 4) as an accumulation of these two [2, 6]; but right at those moments, the induced component is so light and instantaneously last to be really accounted.

d) End-to-End Throughput Theoretical Model

The following graph in (Figure 4) shows the theoretical performance of WLAN or mobile wireless network. The performance degradation based congestions does not reach the core networks. From the literatures, congestion can be broadly viewed as the networks performance degradation origin. Another most important lesson is the great impact on the E2E caused by congestions relatively to the network loads increase (Figure 4). This remark shows that the good design and management of local network hooked to the internet would contribute to networks high performance.

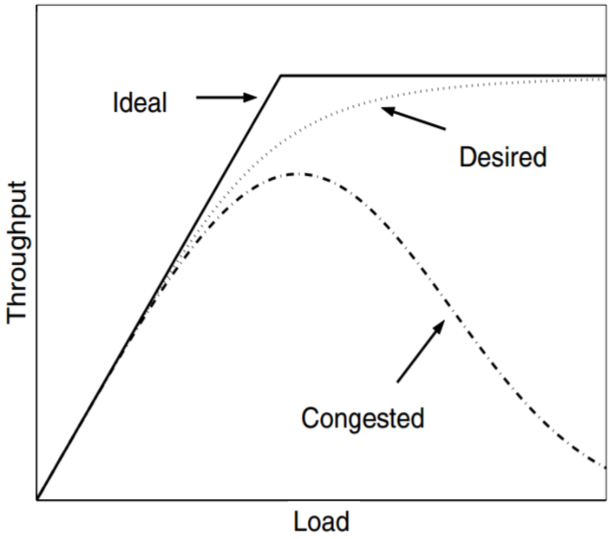


Fig. 4. WLAN theoretical performance graph showing the degradation (E2E decrease) during congested periods [6]

Figure 5 displays the theoretical curve for network congestion in terms of the main influencing factor – network load.

C. Congestion Control Categories

With reference to Figure 4, the following diagram in Figure 5 shows the congestion controls commonly applied methods. In findings, all the reviewed papers showed that congestion happens actually at the open networks level. It occurs typically between different subnets as a result of a probable poor control at LAN/WLAN management level. Therefore, starting at local network’s clients stage, RTS/CTS have a considerable role to play along with other associated tools for performance management at networks level.

Here are some practical solution methods to handle induced and normal congestion.

- RTS/CTS induced congestion can be cut-down by controlling and manually modifying the RTS applied threshold (e.g. packets size) [2, 3. 4. 6]. In fact, various new features supporting RTS/CTS operations make it possible to sharply minimize induced congestion occurrence —e.g. MACAW, which include an ACK at the WLAN MAC level [6].

- LAN/WLAN normal or systematic congestions can expand beyond subnet’s section via the interconnection-points and cause the open networks congestion. However, a set of mechanisms is available for this level of congestion management. For examples Detection and avoidance; control detection and removal when already occurred [22].

Figure 5 displays the congestion control categories and their management relevant policy tools. The implementation decision depends on the problem model and particularly the solution type and scheme (i.e. simple/single mechanism; multiple mechanisms and thus integrated solutions).

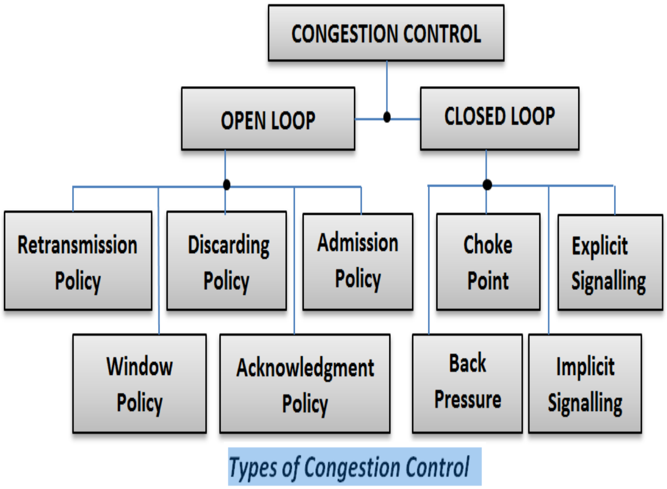


Fig. 5. Congestion controls Common Methods -- the two congestion control categories and their respective policy members – Adapted from [22]

III. Performance and QoS Background of Wired and Wireless Networks

In performance and QoS survey papers, WLAN operation faces mainly multiple technical challenges [12, 22] but, just a huge of efforts are deployed on typical issues over time by the industries and vendors for solutions support or with new systems/products' added features. However, QoS by definition is expected to primarily target user's satisfaction. This would include other satisfaction's aspects like services cost, user (data and personal info) security, mobility and network availability whenever needed, etc. Most of these factors are engineering based tasks. Therefore, the most important can be those allowing users to enjoy the networks use. That will be then a result of a good job done by remote and local solutions support. In general (if not in most cases) users need some immediate and friendly methods/tools (e.g. Simple and direct troubleshooting guides on common surfing issues) for quick help; and that would be good enough to their satisfaction.

What are the listed performance/QoS issues and solution methods in recent survey papers including [11, 23, 24, 31] articles (and unread ones)? And then, what have been proposed for user's emergency basic tools? In findings from reviewed ones, there is little in offer (explanations); whereas much is being said also but, more are in technical ways. And at practical level, possible helpful details (settings) are available only on expensive products (e.g. routers and wireless router (AP) and some end-nodes - workstation, laptops and smart phones among shared systems). Therefore, only WN/WLAN managements generally are able to take required actions when a problem arises. The approach of this papers review is more about the absence of such details in recent study's discussions, which can be a valuable input, a reminder to the networks people on the matter.

- *Performance and QoS Meaning Confusion Impact*

In computing, the word 'performance' has two interpretations: (a) --a computer operation's speed by counting operations or instructions executed, (b) –a computer system outcome in term of "throughput" (i.e. E2E packets # sent or received), node's response time, and availability [25]. Quality has to do with a standard of something considered against (or relatively to) many others of same kind based on its degree of excellence. Such standard is actually hard in wireless network or technology things but achieved generally by relativity to the nearer consensus of people. According to Margaret Rouse (2006) [26], "in information technology product or service, quality is sometimes defined as meeting the requirements of the customer". Thus, the networks QoS focuses on user's opinion (satisfaction) for such standard definition.

Performance and QoS are then better understood mainly from their practical interpretations at the end-users level, which are the LAN/WLAN's clients where application services' outcomes are visualized and thus appreciated. A comprehensible demo with grade of services (GoS) versus QoS is available in the [27] article. Similarly, the structure of a QoS operation from the network to the client level shows that the process is basically about a coordinating activity from different network sections out of which the outcome can be displayed at every LAN/WN's client.

WLAN QoS faces some major technical challenges as illustrated in [28, 29]. Based on the literatures including these above references and particularly [27], network user's experience is much influenced by the cost and marketing of the services (i.e. NSP). Thus, their satisfaction from received services is a mix of non-technical and technical facts. However, their agreement for the QoS relies particularly on the technical result experienced on their terminal (regardless of what has been said much in marketing or selling prices). Their quality of experience (QoE) encompasses their expectation) and facility or system use. Thus, any possible discrepancy or mismatches in their hopes can merely degenerate into discomfort and poor feelings about the QoS [27]. Therefore, if all the great technical works are done for the QoS at the network layer only, it is likely to not actually reach the main targeted objective, which is the customers' satisfaction. They can be offered more friendly use tools to face WN's common issues whenever necessary during networks access. For learning and guidance to solution designers, [29] article introduced about some necessary understanding of the QoS mechanisms as defined today in the IEEE 802.11; and unfortunately, RTS/CTS was not addresses into that well-summarized materials.

IV. WLAN Performance Using RTS/CTS Framework

In general, there are two optional transmission control methods within WLAN environment—the use of physical carrier sensing and enabling/disabling RTS/CTS mechanism [4, 20] for a purpose of performance planning.

A. WLAN MAC Important Functions Overview

Referring to MAC protocols, there are two particular functions configurable cumulatively with RTS frames operation depending on the application services traffic to manage. These are distributed coordination function (DCF) and point coordination function (PCF) [4, 20, 18, 21]. They can be assisted by other sub-system's functions such as. queuing disciplines (QD), enhanced distributed channel access (EDCA); and hybrid controlled channel access / hybrid coordinated function (HCCA / HCF) Their configurations along with RTS/CTS mechanism can help deal with the issue of resource contention between WLAN communicating stations [42]. Figure 6 shows how their functional operation compares and complement each other.

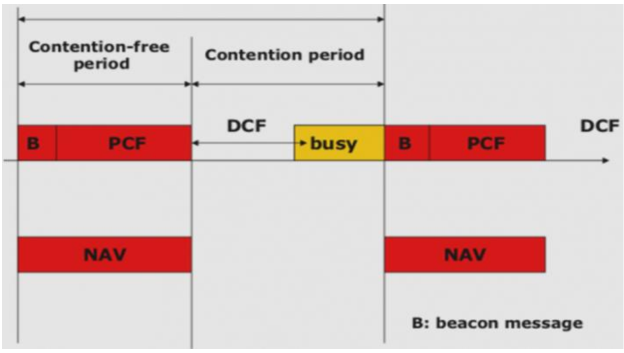


Fig. 6. Comparing DCF vs. PCF operation

According to [16], DCF (if implemented alone) will lead to many collisions at peak periods. Thus, DCF and PCF are genereally configured together for any of applicable schemes in the framework. However, PCF is seen as more useful than DCF; it assists DCF and it particularly enables provisioning WLAN QoS under IEEE 802.11e standard. In fact, PCF let create a model of QoS solution convenient for real-time multimedia applications. Also, [13] noticed that most of designed protocols to overcome hidden and exposed node problems made use of DCF in turn supported by RTS/CTS mechanism.

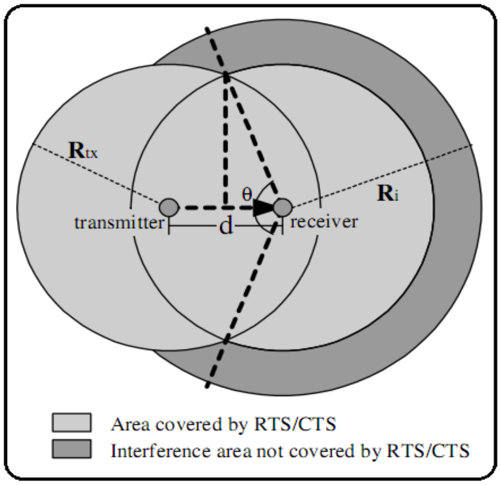

Rtx
Ri
θ
transmitter
d
receiver
▭ Area covered by RTS/CTS
▬ Interference area not covered by RTS/CTS

Fig. 7. RTS/CTS handshake effectiveness estimation —i.e. for (‘d’) larger than 0.56*R_{tx} and smaller than R_{tx} [4]

B. RTS/CTS Framework Mathematical Modeling/Design

Like any other technologies, IEEE 802.11’s RTS/CTS handshake got some limitations as compared to its theoretically expected performance [4].

For instance, this mechanism is not able to fully eliminate the hidden terminal problems. Anyhow, as example, here is an introduction to a mathematical modeling related to some direct parameters that are linked to this complex problem regarding WLAN performance.

1) Modeling the space between hidden nodes for communications

The model of problem on hidden station involved two general elements: the relative location and position of nodes to their local AP and their respective range of transmission power to each other and to the AP position (Figure 7 illustration). All these have in common a distance between the two nodes relatively to an intersection of their (RTS/CTS) radio range coverage compared to the non-covered area.

Based on Figure 7, there are three radio ranges labeled as $R_{tx:}$: transmission range; R_{tx}: carrier sensing range and R_i: interference range. According [3]’s authors, the conditions on the distance “d” to satisfy the receiving node’s signal power (i.e. on the hidden node) must obey the law of the following equation:

$$P_r = P_t G_t G_r \frac{h_t^2 h_r^2}{d^4} \qquad \text{[Eq.1]}$$

with:

- P_t the transmission power; G_t and G_r respectively the antenna gains of transmitter and receiver; h_t and h_r the height of both antennas; d the distance between the transmitter and the receiver.

2) Receiving Station Estimate Signal To Noise Ratio

The demonstration in Figure 7 is with an assumption of being in a homogeneous MANET environment. Therefore a signal arriving at the receiver is considered to be valid if the signal to noise ratio (SNR) is above a certain threshold (SNR_THRESHOLD). Then, SNR is given by SNR=Pr/Pi; and considering homogeneous radios, (SNR) is computed with the equation (2).

$$SNR = P_r / P_i = (\tfrac{r}{d})^4 \geq SNR\_THRESHOLD$$

where $r \geq \sqrt[4]{SNR\_THRESHOLD} * d$ [Eq.2].

This means that to successfully receive a signal, the interfering nodes must be:

$$r \geq \sqrt[4]{SNR\_THRESHOLD} * d \text{ (meters)}$$

away from the receiver. In fact, in practice, SNR_THRESHOLD is usually set to 10. Thus, R_i is as in [Eq.3]:

$$R_i = \sqrt[4]{10} * d = 1.78 * d \qquad \text{[Eq3].}$$

Based on equation (3), when the transmitter- receiver distance “d” is larger than R_{tx}/1.78=0.56*R_{tx} (R_{tx} being the transmission range), the interference range then exceeds the transmission range. This is easy to understand that power level needed for interrupting a transmission is much smaller than that of successfully delivering a packet. The interference area around a receiver is defined as $A_i = \pi R_i^2$. And all the nodes within the interference area will be hidden nodes of the considered receiver.

3) Brief Theory For RTS/CTS Use Effectiveness Planning

Among additional parameters for estimation and prediction in practical control, the effectiveness of RTS/CTS ($E_{RTS/CTS}$) is defined for the following involved elements:

- Ai = Total interference area.

- $A_{iRTS/CTS}$ = Interference area where nodes can receive RTS or CTS successfully.

Hence: $E_{RTS/CTS} = A_{iRTS/CTS} / A_{i.}$ [Eq4].

Then, based on equation (4), for (d <=0.56*R_{tx}), apparently $A_{iRTS/CTS}$ is equal to A_i since transmission range is larger than the interference range. Thus, ERTS/CTS will be almost equal to 1. And when "d" (the transmitter's distance to the receiver's antenna, Fig.7) increases beyond 0.56*R_{tx}, $A_{iRTS/CTS}$ becomes smaller than "A" resulting in the ($E_{RTS/CTS}$) smaller than 1; etc. Further estimation on the RTS/CTS threshold related parameters can be found in [4, 13].

C. Practical Configurations in Performance Design

Here are briefly some commonly recommended RTS/RTS threshold values and range that can be configured in wireless networks testing. In finding, a WLAN permanent monitoring is the only better way to find out which "exact" values to make use as a result of good understanding the network behaviour over some specific periods of time (e.g. day times, week days, etc.). In other words, it is generally recommended to determine appropriate periods when RTS/CTS must be enabled or disabled with suitable threshold settings. For, this added touch is practically the easier way to adjust the WLAN topology change due to users' relative moving position within its wireless coverage environment,

According to [5, 20] there is a range of RTS/CTS (and fragmentation) threshold setting values that can help network manager to choose from after a routine assessment of the network behaviors. The typical activities for carrying out such assessments are discussed in [5, 33}]. As threshold setting examples:

a) When having many users far from the access point, lower the threshold to 2304 bytes; then verify the new outcome;

b) For Fragmentation as solution; default size threshold is 2346 bytes and the standard range is 256-2346 bytes;

c) In real-world, these indicated setting values should be tried between 256 and 2346 until getting the fine tune with the data flow is normalized.

d) Etc.

These above cited examples are among the most common settings in work testing cases from many literatures.

V. RTS/CTS in Integrated Schemes based WLAN Performance Solutions

RST/CTS mechanism is still implemented nowadays in temporary use manner (i.e. manual configuration vs. automated insertion). In fact, from traditional function, it is considerable as associated tools with any QoS solutions implemented to support the network performance. An observation from most of the research works on this mechanism is much more considered as technique most appropriate to combat transmission collisions at wireless AP [2, 3, 4, 6].

As discussed under above section 3, it is important to have a clear understanding between performance and service quality performance. That is because it helps know and localize where their measurement's parameters intervene/act with respect to the network structural layers. This includes the actual role played by RTS/CTS in the assessment of the two, which are commonly ignored or invisible part of the networks integrated solutions. A comprehensible study on IEEE 802.11 WN/WLAN has been carried out in [30, 31] articles about its background, technology standards and applications.

Similar study has been found in related articles reviews and surveys conducted by [28, 24, 43, 44] authors particularly about its (IEEE 802.11) implementation for networks QoS support. For instance, QoS matter is reviewed in some contemporary research articles like [24] with a great attempt of classifying and categorization methods, protocols and methods for network performance and QoS design. Nevertheless, those materials are well-presented, but for the most knowledge people; since all technical details focused only on the latest knowledge or terminologies beyond RTS/CTS in argumentation throughout each articles review. However, another article [45] is an overview of QoS in wireless data networks, which summarized the commonly used tools and technologies, and included the potentials and important role of RTS/CTS for this purpose in WLAN, Hence, the subject of concern in this article is obviously verifiable from such recent survey papers; thanks for these articles contents' quality and coverage.

Furthermore, recent and oncoming trends in networking data communications is the highest interest in using integrated data or multimedia contents, which requires the implementation of QoS integrated solutions (QoS-IS). Based on (Arindam Paul, 1999) [32] 'QoS-IS' is a standards set from Internet Engineering Task Force (IETF) group to support various network traffic classes with different QoS profiles through some network elements. The system generally works fine subject to the resources availability managed by an admissions control system — a switch or router's policy decisions base. Typically, such network system based QoS-IS (provisioning) will involve either or most of the following scheme elements: congestion avoidance/congestion management mechanisms**,** per-flow-state maintenance tools, traffic shaping and policing; and link efficiency control. Hence the lack of this disposition leads to offering all existing resources to any traffic classes and thus leading the network traffic onto a best-effort support. Some of related activities for QoS–IS are illustrated in above Figure 5.

A. RTS Importance and WLAN QoS Provisioning Techniques

The foremost use – enabling RTS operation, is to combat any possible packets collision between clients, which have become invisible or insensible one to another. This happens due to the interferences phenomenon, or clients unreachable in their radio range. Such a situation occurs when user's clients are too much wide spread and thus become unheard in the AP wireless medium coverage.

In overall above situations, RTS/CTS are proven capable in WLAN performance support by testing, then minimizing and avoiding collisions [5, 35, 38]. In RTS/CTS operation (Figures 1, 2 & 3) the station initiating a communication process sends RTS frame to the AP. RTS and CTS exchange acts as environment free testing process. This in turn enables reducing packets transmission collision. And when properly enabled congestion context and threshold settings correct choice), then collisions can sufficiently be avoided [2, 5, 15]. RTS/CTS can also be considered as a fundamental tool for WLAN performance management [23, 36, 41]. That is because the protocols can successfully handle following two critical issues in performance management: (a) Problem of hidden stations [43]; (b) securing performance troubles with extra-protection that reduces/eliminates the risks of collisions [6]. With this way the delay is minimized and the throughput E2E is guaranteed with less data loss. Overall, RTS/CTS offer concrete implementation that assures high probability of performance degradation avoidance; for, it is linked to above (a&b) situations [43] than contrary solutions.

Many research studies have discussed WLAN performance issues along with various enhancement solutions. Contemporary trends in networking applications are likely more about mobile networks deployment and multimedia applications as network contents. Meanwhile, this category of application services are much demanding in their service quality performance requirements which are very sensible to collisions phenomenon. And collisions are proven to have severe and intolerable effects on these popular WLAN/WMN service applications (e.g. voice, audio and rich media) [3, 6, 32, 43] for either of the following reason:

a) Collisions will cause packets loss and thus throughputs decrease.

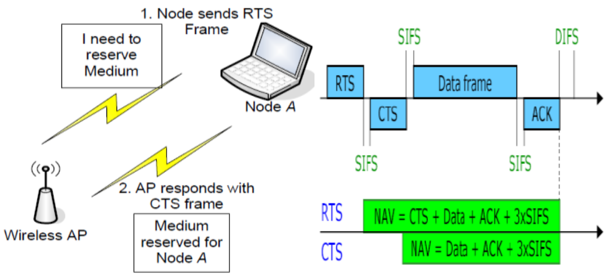
b) Collision will also introduce additional transmission delays other than the systematic one.

At the overall networks level, (a) and (b) cause congestion between LANs' sections, due to missing frames retransmissions, which introduce in most cases an unrealistic congestion [31].

Other additional performance solutions (beyond the scope of this articles review) are generally used in combination with default RST/CTS for network performance enhancement.

A WLAN performance can be improved after an observation of some persisting decreases in throughputs delivery or excessive E2E delays. Relatively to WLAN performance issues, such facts are the revealing effects of collisions between wireless communicating nodes' transmitted packets [1-6, 13, 14, 15]. The additional collision's defects include the data loss and congestions at network level that contribute directly to the performance and QoS degradation.

As simpler and practical solution to these issues, one can turn on (or off) the RTS/CTS protocols on every WLAN clients [2, 6]; they are recognized as powerful for controlling and minimizing collisions happening [2, 3, 17, 43]. Regarding some of RTS/CTS limitations, [15] led a valuable study on some perspectives and came up with some proposed solutions. Different RTS/CTS schemes as framework exist for this purpose (e.g. Bandwidth reservation, reducing delays and loss). Figure 8, refers to the case for BSS based WLAN.


1. Node sends RTS Frame
I need to reserve Medium
Node A
Wireless AP
2. AP responds with CTS frame
Medium reserved for Node A
SIFS
DIFS
RTS
Data frame
CTS
ACK
SIFS
SIFS
RTS
CTS
NAV = CTS + Data + ACK + 3xSIFS
NAV = Data + ACK + 3xSIFS

(1) Request for Bandwidth Reservation based

(2) Handshake associated Timeline

Fig. 8. Bandwidth Reservation request using RTC/CTS mechanism

In this case, the RTS/CTS scheme consists of five frames (Fig. 8(2)); but the bandwidth reservation is ensured actually by two: Short Inter-frame Space (SIFS) and Network Allocation Vector (NAV) [2]. This framework is associated with the called single exclusive access or shared MAC [3, 6, 15] configurations mode.

However, with distributive MAC (DMAC) known also as multiple accesses MAC, the RTS/CTS framework differs significantly according to the wireless network (WN) deployment's access schemes in use. That is whether it applies a random access or a controlled access techniques [18] to suit the WN deployment. The most important detail is that, DMAC is the scheme appropriate for MANET environment [6, 13, 15, 18] in order to provide ad-hoc wireless mobile node's services application with an acceptable QoS level. QoS problems get more complex due to different factors. These include an attempt to accommodate various application services concurrently running (even for those without any performance requirements). Another factors case and the most challenging is about keeping healthy the network state information accuracy, which is merely void [27]. And, this impossibility has a room for a satisfactory solution under WLAN/WN managers' duty; that is about well-monitoring WLAN and using wisely RTS/CTS features in order to prevent or at least minimize the congestion occurrence known up to here as the bottom/root cause of the network poor performance and then QoS degradation.

VI. Conclusion

This paper has discussed about RTS/CTS framework paradigm, and particularly the important role that this mechanism plays along with existing WLAN QoS provisioning methods. A review on heterogeneous networks (e.g. taking multimedia network as a general model) using both early and recent literatures has proven that networks congestion can be considered as the major root for networks performance degradation. In fact, in wireless networks environment, packets transmission collision is technically the primary source of data loss; E2E decrease is the immediate effect. However, RTS/CTS stand as the fundamental mechanism that is suitable and simpler solution to the above discussed WLAN systematic issue. Thanks to WLAN and network technologies evolution, this tool is friendly made accessible on some WLAN routers.

A particular contribution of this review has been to demonstrate that these basic and simpler tools are kept mute (absent) in most study papers that discussed about WLAN performance / QoS issues. Even though RTS/CTS alone may not provide today's WN model with sufficient QoS support, they remain among the most powerful tools for WLAN managers' support. For, they help those managers in allowing WLAN clients to enjoy typical QoS (if any) as offered by their services provider. Moreover, in addition to the source origin of networks congestion and the growing trends of service applications, new QoS factors are added up. Thus, the following details show that we will not get rid of their causes for soon. They are the technology with endless imperfections and limitations, user's increasing demands for services higher quality/features; etc. Therefore, the end of the stated issues needs to be always considered with their fundamental solution's tools (RTS/CTS) as far as networking communications will still be alive and using IEEE 802.11.

A future work for this articles review will be a practical lab testing on the efficiency of the RTS/CTS mechanism features as QoS support for multimedia based mobile WLANs networks; including their direct insertion into multiple schemes based QoS integrated solutions.

ACKNOWLEDGMENT

The authors would like to thank the Universiti Teknikal Malaysia (UTeM) through its research and innovations department services (CRIM) for the payment of this journal article registration and publication fees (Footnote extension from first page).

REFERENCES

[1] Wi-fiplanet.com, Jim Geier, 2002 Improving WLAN Performance with RTS/CTS, Wi-Fi Planet Tutorials; accessed: 25-11-2016, from [http://www.wi-fiplanet.com/tutorials/article.php/1445641/Improving-WLAN-Performance-with-RTSCTS.htm].

[2] Shakil Akhtar, 2006 Communication Performance of 802.11 WLANs; Proceedings of the 10th WSEAS International Conference on APPLIED MATHEMATICS, Dallas, Texas, USA, November 1-3, 2006 89; accessed: 29-11-2016, from [http://www.wseas.us/e-library/conferences/2006dallas/papers/519-256.pdf].

[3] Hetal Jasani & Nasser Alaraje, 2007 Evaluating the Performance of IEEE 802.11 Network using RTS/CTS Mechanism, IEEE EIT 2007 Proceedings; 1-4244-0941-1/07/$25.00 c 2007 IEEE; retrieved: 10-03-2016.

[4] Kaixin Xu, Mario Gerla & Sang Bae, n.d. How effectiveness is RTS/CTS handshake in IEEE 802.11 based ad hoc networks? University of California, Los Angeles, Computer Science Department, Los Angeles, CA 90095, USA; retrieved: 10-03-2016.

[5] INFOSEC, 2014 RTS threshold configuration for improved wireless network performance; accessed: 10-03-2016, at: [http://resources.infosecinstitute.com/rts-threshold-configuration-improved-wireless-network-performance/].

[6] Saikat Ray, Jeffrey B. Carruthers, and David Starobinski, n.d. RTS/CTS-induced congestion in ad hoc wireless LANs (Lecture); Department of Electrical and Computer Engineering --Boston University; accessed: 10-03-2016, at: [www.cacs.louisiana.edu/.../fas9529.p...].

[7] Jaime Lloret, Miguel Garcia, Hugo Coll and Miguel Edo, 2012 Wireless sensor networks and systems --Wireless Technologies: Concepts, Methodologies, Tools and Applications, DOI: 10.4018/978-1-61350-101-6.ch102; Pages: 13; Copyright: © 2012; retrieved: 29-11-16.

[8] Wireless local area network (WLAN) gateway system; accessed: 22-22-10-2016, at: [http://www.google.com/patents/US20130103558].

[9] Lingnan University, 2009 Background and History of Wireless LAN for Lingnan University; Copyright© 2016 Lingnan University All rights reserved; accessed: 09-09-16, from [https://www.ln.edu.hk/itsc/network/wireless/bg].

[10] Chaim Ziegler, 2009 Wireless LAN Applications, Wireless; retrieved: 11-11-2016; Copyright © 2009.

[11] Kensuke Miyashita & ,Yuki Maruno, 2016 Campus Wireless LAN Usage Analysis and Its Applications; Chapter Neural Information Processing; Volume 9947 of the series Lecture Notes in Computer Science pp 563-569; Springer International Publishing; DOI: 10.1007/978-3-319-46687-3_62; Print ISBN: 978-3-319-46686-6; retrieved: 06-12-2016.

[12] USA Information Resources Management Association, 2012 Wireless Technologies: Concepts, Methodologies, Tools and Applications (3 Volumes) Information Resources Management Association (USA) ;| _Copyright: ©2012; Pages: 2358, ISBN13: 9781613501016|ISBN10: 1613501013|EISBN13: 9781613501023; DOI: 10.4018/978-1-61350-101-6; retrieved: 20-11-16.

[13] Jayasuriya, Sylvie Perreau, Arek Dadej, and Steven Gordon, 2004 Hidden vs. exposed terminal problem in ad hoc networks.; Australian telecommunication networks & applications conference (ATNAC) 2004 (2004) pp. 52-59; retrieved: 22-07-16.

[14] Khushboo Agarwal and Vikas Sejwar, 2015 Avoidance of hidden terminal & exposed terminal problem using directional MAC protocol; International Journal of Future Generation Communication and Networking ; Vol. 8, No. 4 (2015), pp. 231-238; ISSN: 2233-7857 IJFGCN Copyright ⓒ 2015 SERSC; retrieved: 20-07-16.

[15] L. Boroumand, R. H. Khokhar, L. A. Bakhtiar and M. Pourvahab, 2012 A review of techniques to resolve the hidden node problem in wireless networks. Smart Computing Review, vol. 2, no. 2, April 2012; DOI: 10.6029/smartcr.2012.02.001; retrieved: 20-07-16.

[16] Tankonyvtar.Hu/En, Vilmos Simon , 2014 Wireless and mobile technologies for the future internet, article online; accessed: 12-07-16, at [http://www.tankonyvtar.hu/en/tartalom/tamop412A/20110050_09_wireless_mobile_technologies/ar01s03.html].

[17] www.comlab.hut.fi, Teknillinen Korke Akoulu, n.d. Lecture 4: WLAN 2 (MAC layer operation); accessed: 12-07-16, at: [www.comlab.hut.fi/studies/3240/luentokalvot/4_wlan2.ppt].

[18] Andrew Von Nagy , 2011 Understanding Wi-Fi carrier sense; accessed: 28-06-16, at: [http://www.revolutionwifi.net/revolutionwifi/2011/03/understanding-wi-fi-carrier-sense.html].

[19] www.cs.jhu.edu, David Holmer, 2002. Wireless medium access, lecture; retrieved: 12-12-16.

[20] INFOSEC, 2014 RTS threshold configuration for improved wireless network performance; accessed: 28-06-16, at: [http://resources.infosecinstitute.com/rts-threshold-configuration-improved-wireless-network-performance/].

[21] Lecture, Markku Renfors, n.d. IEEE 802.11 ; (IEEE 802.11 overview, architecture and MAC); accessed: 28-06-16, at: [http://docplayer.net/12481121-802-11-markku-renfors-partly-based-on-student-presentation-by-lukasz-kondrad-tomasz-augustynowicz-jaroslaw-lacki-jakub-jakubiak.html].

[22] E-Computernotes.com, Dinesh Thakur, n.d. What is congestion control? Describe the congestion control algorithm commonly use; accessed: 28-06-16, at: Available at: [ecomputernotes.com/computernetworkingnotes/...networks/].

[23] Kensuke Miyashita & ,Yuki Maruno, 2016 Campus Wireless LAN Usage Analysis and Its Applications; Chapter Neural Information Processing; Volume 9947 of the series Lecture Notes in Computer Science pp.563-569; Springer International Publishing; DOI: 10.1007/978-3-319-46687-3_62; Print ISBN: 978-3-319-46686-6; retrieved: 06-12-2016.

[24] Qiang Ni, Lamia Romdhani and Thierry Turletti, 2004 A survey of QoS enhancements for IEEE 802.11 wireless LAN: research articles; Journal Wireless Communications & Mobile Computing archive; Volume 4 Issue 5, August 2004, Pages 547 - 566 , DOI: 10.1002/wcm.v4:5; John Wiley and Sons Ltd. Chichester, UK; retrieved: 06-12-2016.

[25] Whatis.techtarget, Margaret Rouse, 2006 What is performance?; Part of the Computing fundamentals glossary: Definition from WhatIs.com; accessed: 27-11-2016, from [http://whatis.techtarget.com/definition/performance].

[26] Whatis.techtarget.com, Margaret Rouse2006 What is quality? Part of the Programming glossary: Definition from WhatIs.com; accessed: 27-11-2016, from [http://whatis.techtarget.com/definition/quality].

[27] Amandeep Kaur, 2011 An overview of quality of service computer network; Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 3 Jun-Jul 2011; ISSN: 0976-5166; retrieved: 10-11-2016.

[28] Richard Hill, 2012 Overview of quality of service (QoS); APT-ITU workshop on the International Telecommunica3ons Regulations Bangkok, 6-8 February 2012 Richard Hill, ITU; retrieved: 20-10-2016,

[29] BROADCOM, Philippe Klein, 2008 802.11 QoS overview; IEEE Plenary Meeting – Nov 08 Dallas, TX, avb-phkl-802-11-qos-overview-0811-1; retrieved: 20-10-2016.

[30] Ibrahim Al Shourbaji, (n.d.) An Overview of Wireless Local Area Networks (WLAN); Computer Networks Department; Jazan University; Jazan 82822-6649; retrieved: 09-09-2016**.**

[31] Kevin J. Negus, and Al Petrick, 2008 History of Wireless Local Area Networks (WLANs) in the Unlicensed Bands ; George Mason University Law School Conference, Information Economy Project, Arlington, VA., info, Vol. 11 Iss: 5, pp.36 – 56; DOI: http://dx.doi.org/10.1108/14636690910989324; 09-09-2016.

[32] Arindam Paul, 1999 QoS in Data Networks: Protocols and Standards; accessed: 28-11-2016, from [http://www.cse.wustl.edu/~jain/cis788-99/ftp/qos_protocols.pdf].

[33] *www.ciscopress.com, 2010* Performance Considerations - WLAN Design: Range, Performance of WLAN; accessed: 18-05-16.; at: [www.ciscopress.com/articles/article.asp?p=1613796...3].

[34] Home-Network-Help.Com,*n.d.* Improving wireless network performance by tuning advanced wireless settings on wireless device; accessed: 18-05-16.; at [http://www.home-network-help.com/wireless-network-performance.html].

[35] Wang, Kaishun Wu and Mounir Hamdi, 2012 Combating hidden and exposed terminal problems in wireless networks; IEEE Transactions on Wireless Communication, vol.10, no.10, 40-12, 2012; retrieved: 25-11-16.

[36] Hui Ma, Eiman Alotaibi, and Sumit Roy, n.d. Analysis and simulation model of physical carrier sensing in IEEE 802.11 mesh networks; retrieved: 25-11-16.

[37] Felix Diaconu, 2012 IEEE 802.11 MAC frame fragmentation performances in jammed environments**;** Bul. Inst. Polit. Iaşi, t. LVIII (LXII), f. 2, 2012l; retrieved: 17-06-16.

[38] Jaime Lloret, Miguel Garcia, Hugo Coll and Miguel Edo, 2012 Wireless sensor networks and systems --Wireless Technologies: Concepts, Methodologies, Tools and Applications, DOI: 10.4018/978-1-61350-101-6.ch102; Pages: 13; Copyright: © 2012. Retrieved: 29-11-16.

[39] Stefan Mangold, Sunghyun Choi, Peter May, Ole Klein, Guido Hiertz & Lothar Stibor, n.d. IEEE 802.11e wireless LAN for quality of service. Retrieved: 29-11-16.

[40] Mangold, Sunghyun Choi, Guido R. Hiertz, Ole Klein,& Bernhard Walke, 2003 Analysis of IEEE 802.11e for QoS support in wireless LANs*;* IEEE wireless communications, volume:10 , issue: 6, page(s): 40 – 50; ISSN : 1536-1284 ; INSPEC accession number: 7873965; DOI: 10.1109/MWC.2003.1265851; IEEE communications society. retrieved: 17-06-16.

[41] Namita Yadav & Sanjay Sachan, 2014 Analysis and the performance effectiveness of RTS /CTS mechanism in IEEE 802.11*;* International Journal of Emerging Trends & Technologies in Computer Science (IJETTCS); Volume 3, Issue 4, July-August 2014, ISSN 2278-6856; retrieved: 18-05-16.

[42] Ashwini Dalvi, Pamukumar Swamy and B B Meshram, 2011 DCF improvement for satisfactory throughput of 802.11 WLAN; International Journal on Computer Science and Engineering (IJCSE);Vol. 3 No. 7 July 2011**;** ISSN : 0975-3397; retrieved: 18-05-16.

[43] Journal of Engineering Research & Technology; Volume/Issue: Vol.2 - Issue 4 (April - 2013), e-ISSN: 2278-0181; retrieved: 18-05-16.

[44] Hua Zhu, Ming Li, Imrich Chlamtac, and B. Prabhakaran. 2004, A survey of quality of service in IEEE 802.11 networks; IEEE Wireless Communications, volume: 11, issue: 4 ; page(s): 6 - 14 ; ISSN : 1536-1284; INSPEC Accession Number: 8086234; DOI: 10.1109/MWC.2004.1325887; IEEE communications society; retrieved: 18-05-16.

[45] Adam Petcher, n.d. QoS in Wireless data networks (Online article); accessed: 11-07-16.; at [http://www.cs.wustl.edu/~jain/cse574-06/ftp/wireless_qos/].

Finite Element Method Combined with Neural Networks for Power System Grounding Investigation

Liviu Neamt, Oliviu Matei, Olivian Chiver
Electrical, Electronic and Computer Engineering Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

***Abstract*—Even in homogenous soil and for simple geometrical structure the analytical design of a grounding system is a complex and not very accurate procedure. Using Finite Element Analysis (FEA) it can perform a precise design for complex grounding systems but with important hardware resources and time consumption. This paper proposes a methodology for power system grounding design, directed to ensure the advantages of the FEA but without its disadvantages. This is realized by adding the function emulation using neural networks. The vertical rod, buried in inhomogeneous soil is the subject of this presentation. Consequently, the first step was to perform FEA for a large number of configurations: different types of vertical rods connected to the surface, buried at various depths in different double-layer soil structures. Then, the results have been interpreted through a multi-layer perceptron (MLP) with one hidden layer. A compromise between the number of inputs and precision have been tested, in order to define a minimum number of FEA required to obtain an acceptable grounding system design, i.e. a desired grounding resistance, for any combinations of the geometrical and material parameters. The validation of the methodology was done based on data reported in various research works.**

Keywords—neural network; finite element analysis; power systems; grounding

I. Introduction

Numerical simulation of the electromagnetic field lays on the basis of modern CAD in electrical engineering. Finite Element Method (FEM) is the most used tool for this and permits, not just the design, but also the optimization and the validation of the equipment behavior in the field. The main disadvantage consists in hardware resources and time needed for simulations and the lack of generalization, i.e. for every configuration it must be performed another FEA

The targeted sustainable smart grid concept, which ensures the continuity and quality of the energy supplies, must be realized in order to guarantee the safety of the human being and the installations. One of the key for achieving this is the power system grounding.

The professional design of power system grounding is conducted analytical in homogenous soil, for simple structures configurations, according to the theoretical computation of the electromagnetic field [1], [2] and again for more complex devices, but using simplified relations imposed by standards and regulations, [3-5]. For inhomogeneous soils, analytical relations are very complicate, if exist, inaccurate and determinable for simplified structures, material parameters and variable behavior [1-5]. All these difficulties can be easily hurdle using FEA [6-11]. The problem here is that the results cannot be generalized, so for every configuration it means another simulation.

This paper intend to structure a methodology, based on FEA and neural network to generalize the FEA result, meaning the grounding resistance value, for any variation of the geometrical and material parameter of the base configuration. For an easy and logic presentation, the methodology is depicted using a usable grounding structure, i.e. a single vertical rod, with variable length, buried in double layer, horizontally layered soil and connected to the surface.

Therefore, an initial configuration, changed successively regarding the imposed variation limits for parameters (geometrical and material) will constitute the FEA models. In this step, a large number of models will be analyzed, asking for great hardware and time resources.

All the results, i.e. the grounding resistances, in terms of: rod length, the thickness of the first soil layer and the ratio between upper and inferior layer resistivity, will enter in the next stage, meaning the neural network generalization. Creation, optimization of the neural network, reducing the number of inputs required to maintain a desired precision are the goals of this last step.

As final result, for the above grounding system, the methodology offers, virtually instantaneously, a value of the grounding resistance for any combination of the material and geometrical parameters, offering in this way a key to a close optimum configuration.

II. Vertical Rod Grounding Structure

Standard regulations provide minimum allowed cross section for different materials for electrodes and also suggest some recommendations for different configurations. Based on these, the basic configuration is a Zn coated steel cylindrical rod, length L = 1 ÷ 3 m, diameter d = 2 inches, buried at a depth h = 1 m, in a double layered soil with resistivity, ρ_{up} and ρ_{inf}, as depicted in Fig. 1.

The FEA for these configurations are presented in details in [12]. The results kept in these step, are the grounding resistances for 300 configurations generated by variations of the next variables: L = 1 ÷ 3 m, the thickness of the upper layer t_{up} = 0 ÷ ∞, the ratio between upper and inferior layer resistivity $\rho_{up}/\rho_{inf} = 0.2 \div 5$.

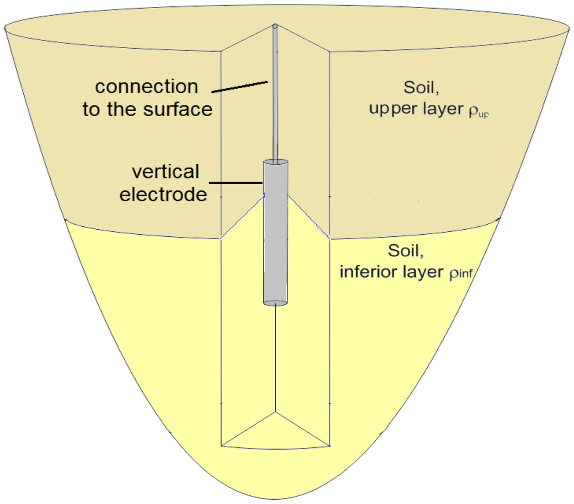


Fig. 1. Vertical grounding electrode with connection to the surface

III. Neural Network Structure

The grounding resistance has been emulated by building a neural network. It is well known the capability of neural networks to approximate functions, a concept called “regression” [13]. For such tasks, a simple multi-layer perceptron has been proved as a good choice, according to Schürmann [13]. However, the complexity of the network is crucial for its behavior. A trade-off regarding its size is always needed. A small architecture may prove inefficient to approximate the desired function, whereas a larger network may over-learn the training set, being unable to generalize on extra input data. Baum and Haussler give some principles in [14].

Therefore several experiments have been run for choosing:

- ✓ The right size of the neural network;
- ✓ The transfer function;
- ✓ The learning algorithm;
- ✓ The training epochs.

The data consisted of 300 sets, with: vertical electrode length, the thickness of the upper layer and the ratio between upper and inferior layer resistivity as inputs for the network and the resistance as the desired output. For avoiding the over-training, which is a common problem is neural design (see [15] and [16]), we have set aside 20% of the data for cross validation and other 20% for testing. This means that the training has been done using only 60% of the available data. For a better learning, the data has been shuffled as suggested by Ikegaya in [17].

A fully-connected multi-layer perceptron has been used, like in Fig. 2.

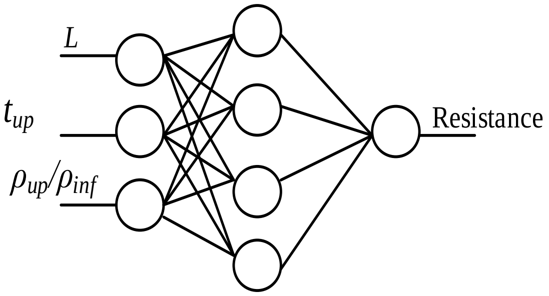


Fig. 2. The general architecture of the neural network tested within the experiments

The neural network has:

- ✓ three inputs, each corresponding to: vertical electrode length, L, the thickness of the upper layer, $t_{up,}$ and the ratio between upper and inferior layer resistivity, ρ_{up}/ρ_{inf};
- ✓ a hidden layer with variable number of neurons;
- ✓ and one output – the desired grounding resistance.

As transfer function we have tested the sigmoid and the hyperbolic tangent. The sigmoid function is defined as:

$$S = \frac{1}{1+e^{-t}} \tag{1}$$

and looks like in Fig. 3.

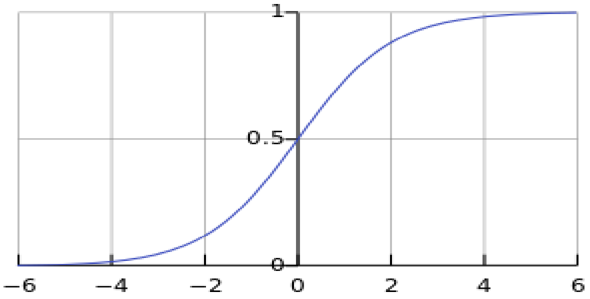


Fig. 3. The shape of the sigmoid function

The hyperbolic tangent (*tanh*) has a shape like in Fig. 4.

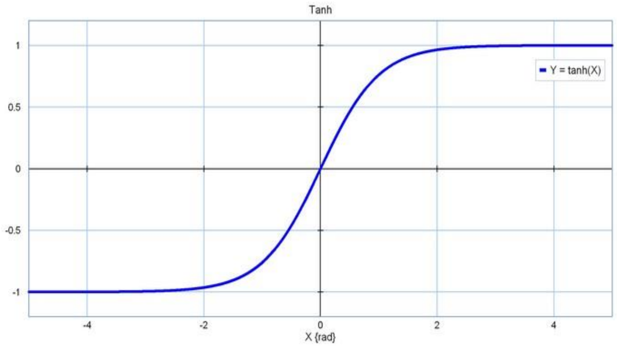


Fig. 4. The shape of *tanh* function

Tanh is defined as:

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{1-e^{-2x}}{1+e^{-2x}} \qquad (2)$$

The learning rules considered during our experiments were: step, momentum, quickprop, delta-bar-delta, conjugate-gradient, Levenberg-Marquardt and resilient backpropagation (rprop), as defined in [18].

Gradient descent learning rules (*Step*) estimate the way to the minimum error of the network. The algorithm searches for the descending slope of the function with various steps. Tweaking the steps is an important aspect of the approach, as smaller steps would result in longer times to reach the optimum, whereas larger steps could overshoot the bottom, causing it to rattle or even diverge.

The *Momentum* provides the gradient descent with some inertia, moving downwards based on some average estimates of that direction.

The *Quickprop* implements Fahlman's quickprop algorithm. It is a gradient search procedure, however very fast in various problems. It is also very accurate. It makes use of the second order derivative for accelerating the search, unlike the *step* or *gradient* methods.

Delta-Bar-Delta is a search method which makes use of the sign of the current update with respect to the previous one. If the two updates are both of the same sign, it increases the learning rate linearly. If the updates have different signs, this is an indication that the weight has been moved too far. When this happens, the learning rate decreases geometrically to avoid divergence.

Conjugate gradient is also a second order method (like *quickprop*), which means that it approximates the second derivatives of the performance surface to determine the weight update.

The *Levenberg-Marquardt* (LM) algorithm is one of the most appropriate higher-order adaptive algorithms known for minimizing the minimum square root (MSE). It is also a second order method. The LM makes use of the so called Gauss-Newton approximation that keeps the Jacobian matrix and discards second order derivatives of the error (see [19]).

Resilient backpropagation (Rprop) is able to outperform most other local (i.e., first-order) learning techniques because it is able to adapt the step sizes of each individual weight instead of using the same step size for all weights. A detailed description of the algorithm can be found at [21].

IV. Researching the Best Neural Architecture

The first experiments were meant to determine the right learning algorithms. The number of hidden neurons was set to 10, and sigmoid as transfer function, as this seems to be a reasonable architecture, according to [12].

The results are concluded in table I. Training MSE are the mean square errors of the neural network in the training phase. The training epochs represents the number of epochs needed for training. This is determined when there are no improvements in the training MSE and before the cross validation starts increasing, as this is the moment when the neural network starts overlearning the training set and behaving poorly on a test set.

“Step”, “Quickprop”, “Deltabar” and “Rprop” present the training MSE, number of training epochs and test MSE when the network uses learning rule: step, quickprop, deltabar, respectively rprop; columns “Mom.”, “Conj. grad.”, “L.M.” show the values when the network is trained with momentum; conjugate gradient, respectively Levenberg-Marquardt.

TABLE I. The Results of the Experiments with 10 Neurons in the Hidden Layer, for Determining the Proper Learning Rule

	Step	**Mom.**	**Quick-prop**	**Delta-bar**	**Conj. grad.**	**L.M.**	**Rprop**
Train-ing MSE	0.47	0.475	0.485	0.275	0.47	**0.266**	**0.266**
Train-ing epochs	142	3000	3000	1720	243	133	**123**
Test MSE	0.50	0.482	0.484	0.363	0.481	**0.345**	0.360

As observed in the table I, the best results, meaning lowest training and test MSE, are obtained using deltabar, Levenberg-Marquardt and rprop as learning rules. The number of training epochs (and thus the learning time) is the best for Levenberg-Marquardt and rprop and significantly higher for deltabar. However, as at this stage of experiments the quality is more important than the time, further experiments have been made taking into account these three algorithms.

We have run (meaning training and testing) the neural network having 5, 10 and 15 neurons in the hidden layer. The number of inputs and outputs remained the same throughout all our experiments.

The results are synthesized in table II. The second column shows the number of neurons in the hidden layer, namely 5, 10 and 15; the next columns presents the training MSE, the number of training epochs and the test MSE when the network uses delta-bar-delta, Levenberg-Marquardt, respectively rprop for learning.

TABLE II. The Results of the Experiments for Determining the Proper Number of Neurons in the Hidden Layer

	No. of hidden neurons	**Deltabar**	**Levenberg-Marquardt**	**RProp**
Training MSE	5	0.2966	**0.273**	0.2787
Training epochs	5	1836	**34**	35
Test MSE	5	0.3543	**0.3457**	0.35191
Training MSE	10	0.2757	**0.2661**	**0.2661**
Training epochs	10	1720	133	**123**
Test MSE	10	0.3631	**0.3455**	0.3607
Training MSE	15	0.2932	0.2674	0.2656
Training epochs	15	1730	52	**48**
Test MSE	15	0.3628	**0.3407**	0.3478

Based on the results shown in table II, a neural network with 5 hidden neurons and L.M as learning rule would be very suited. However, the experiments have shown that it is very unstable, even in the training phase, e.g. the learning curve and the cross validation curve cross each other many times, as shown in Fig. 5.

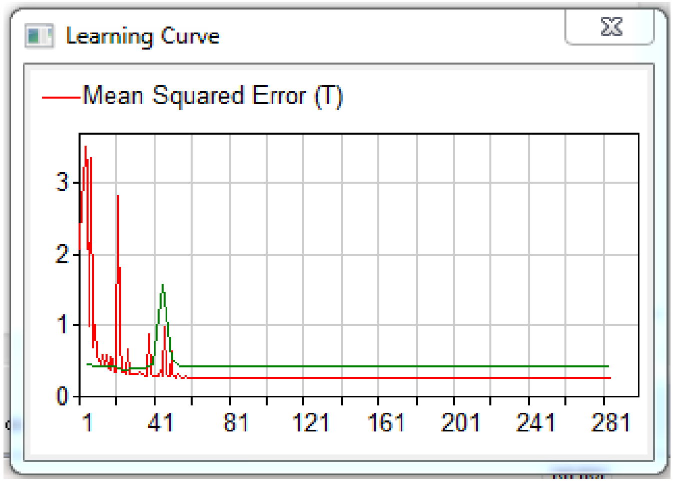


Fig. 5. The MSE for training (red) and cross validation (green) for a neural network with 5 hidden neurons

Therefore so far, the network with 10 hidden neurons and L.M. as the learning rule has the best behavior.

The next experiments were meant to determine the proper transfer function. The same set of tests have been run, but using *tanh* as transfer function, rather than sigmoid. The results are concluded in table III.

The lines represent the training mean square error, respectively training error, and cross validation MSE, respectively cross validation error, the number of training epochs and the testing MSE, respectively testing error. The columns represent the number of neurons in the hidden layer, the results for deltabar, Levenberg-Marquardt and rprop learning rules.

TABLE III. THE RESULTS OF THE EXPERIMENTS USING *TANH* AS TRANSFER FUNCTION

	No. of hidden neurons	**Deltabar**	**Levenberg-Marquardt**	**RProp**
Training MSE	5	0.0119	0.01	0.012
Training error	5	13.012%	12.50%	14.25%
CV MSE	5	0.0101	0.005	0.01
CV error	5	14.04%	13.14%	13.48%
Training epochs	5	1622	56	1604
Test MSE	5	0.14	0.002	0.005
Test error	5	79.25%	14.90%	19.23%
Training MSE	10	0.04	0.002	0.002
Training error	10	35%	**7.31%**	8.28%
CV MSE	10	0.02	0.003	0.003
CV error	10	25%	**7.77%**	7.42%
Training epochs	10	414	78	1245
Test MSE	10	0.04	0.0009	0.001
Test error	10	61.30%	**7.88%**	10.85%
Training MSE	15	0.03	0.01	0.0055
Training error	15	36.61%	20.06%	10.96%
CV MSE	15	0.0328	0.01	0.003
CV error	15	34.91%	13.00%	8.56%
Training epochs	15	324	268	1711
Test MSE	15	0.0099	0.01	0.004
Test error	15	29.76%	31.19%	16.88%

The chart in Fig. 6 concludes the number of training epochs for neural networks with 5, 10 and 15 hidden neurons, and sigmoid, respectively *tanh* as transfer functions, for L.M. and Rprop learning rules. The Delta-bar-delta has been skipped as the training times are much higher.

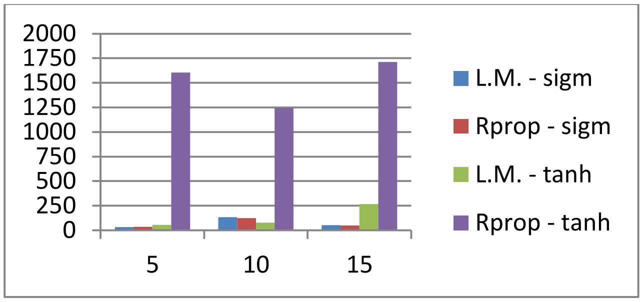


Fig. 6. The number of training epochs for various configurations of the neural network

The training MSE are depicted in Fig. 7.

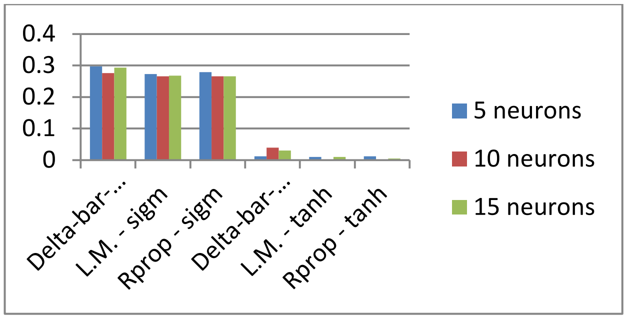


Fig. 7. The training MSE for the researched neural configurations

The MSE of the tests are depicted in Fig. 8

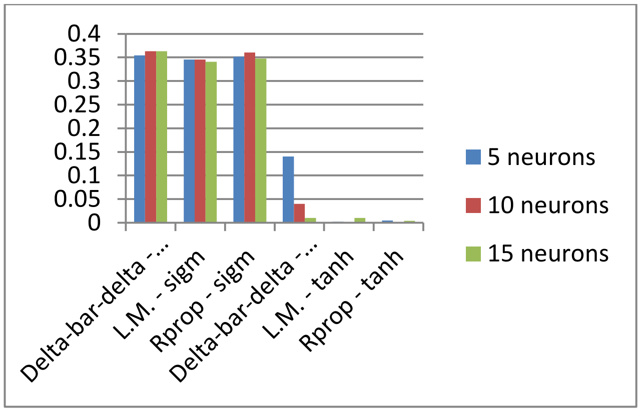


Fig. 8. The testing MSE for the researched neural configurations

From both Fig. 7 and Fig. 8, it is obvious that *tanh* is a better option for the transfer function.

Fig. 9 displays the errors for training, cross-validation and tests. Please notice that there are two training errors for Delta-bar-delta, respectively for 5 and 10 hidden neurons very high (79.25%, respectively 61.30%). But for the sake of clearance, we preferred to scale down to 42% the Y axis of the chart, so that the important errors are visible too.

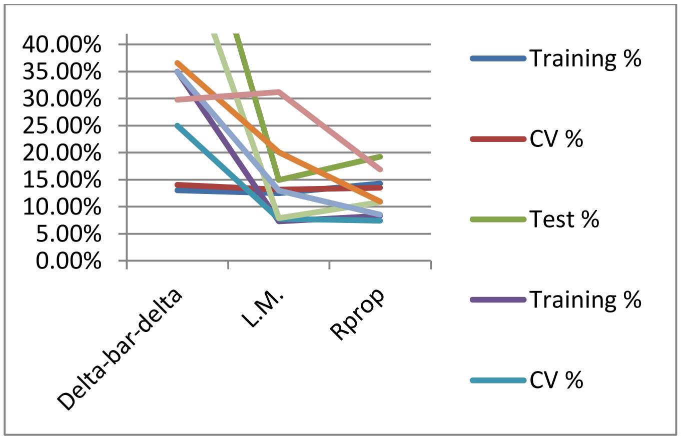


Fig. 9. The training, CV and test errors for the neural configuration with *tanh* as transfer function

V. Experimental Results

At this point we have decided that a neural network with 3 inputs, 10 hidden neurons and one output, using *tanh* as transfer function, respectively L.M. as learning rule is the best option in terms of quality (lowest error), learning time (shortest training time) and computation resources. We have done all these tests using a set of 300 data, out of which 60% (180 records) have been used for training, 20% (60 records) for cross validation and 20% (60 records) for testing. The records belonging to one set or the other have been chosen at random.

The question that rises at this stage is: what is the minimum size of the training set so that the quality of the output is still good (e.g. the error is less than 10%). For that, we have varied the size of the training set from 10% (meaning 30) to 60% (180 records). The training error, cross validation error and test error are depicted in Fig. 10.

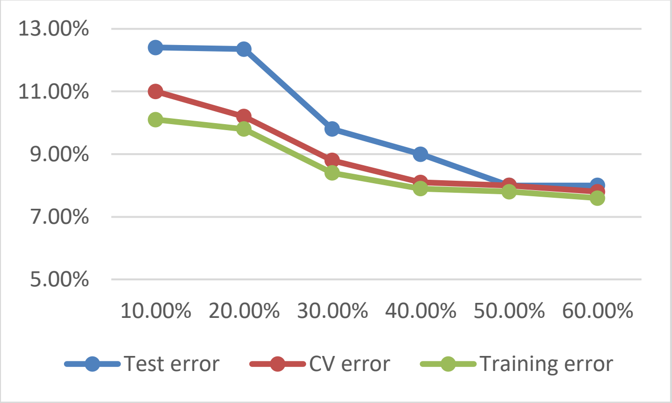


Fig. 10. The errors of the neural network varying the size of the training set

It is obvious that the errors are below 15% in all cases, however, the errors are less than 10% when the size of the training set is above 100 records (30% of the current available data).

This means that 100 FEA must be realized in order to have enough data to be able to generalize the results for all input data in designing process, with a guaranteed error below 10%.

For this grounding system configuration, a value of the grounding resistance for any combination of the material and geometrical parameters is obtained very fast, so the optimization process could be started up. Usually, the soil structure is known and the length of the electrode has to be chosen.

VI. Conclusions

This article presented the experiments and the results obtained for configuring a neural network capable of emulating the grounding resistance of a vertical electrode buried in a two-layered soil.

The presentation was structured on two horizontally layered soil, which for zero or infinite thickness of the upper layer simulate also the homogenous soil. The number or/and the arrangements of the layers could be modified, or/and the configuration of the grounding system could be altered (e.g. horizontally electrode, more simple electrodes, or a complex structure as it is for power substations) so the generality could be pushed further.

Based on methodology depicted above we suggest that FEA combined with neural network analysis may be considered as the best computer aided investigation, not only for power system grounding systems, but also for many other systems used in engineering.

References

[1] O. Centea, Grounding devices from electrical installations, Prizele de pamant din instalatiile electrice, Ed. Academiei, Bucharest, 2006, pp.89-381.

[2] J. He, R Zeng and B Zhang, Methodology and technology for power system grounding, Singapore: J. Wiley &Sons Oxford, 2013, pp.11-19.

[3] IEEE 80-2000 Standard, Guide for safety in ac substation grounding, 2000.

[4] IEEE 142-2007 Standard, Grounding of industrial and commercial power systems, 2007.

[5] 1 RE-Ip 30/2004 Guide book for design and execution of power system grounding, Indreptar de proiectare si executie a instalatiilor de legare la pamant, 2004.

[6] F. P. Dawalibi, J. Ma and R. D. Southey, "Behaviour of Grounding Systems in Multilayer Soils", IEEE Trans. Power Delivery, Vol. 9, No. 1, pp. 334-342, January 1994.

[7] J. A. Guemes and F. E. Hernando, "Method for Calculating the Ground Resistance of Grounding Grids FEM", IEEE Trans. on Power Delivery, vol. 19, no. 2, pp. 595-600, April 2004.

[8] P. Hajebi, A. A. Heidari and A. Mirzaei, "Resistance to Earth of Grounding Grids in Two-layer soil structure using FEM and GA", PIERS Proceedings, Xi'an, China, March 22-26, 2010.

[9] F. Rodriguez, J. A. Guemes, J. M. Ruiz and F. E. Hernando, "Determination of the ground Resistance and Distribution of Potentials in Grounding Grids using FEM", IEEE Transactions On Power Delivery, vol. 21, no 3, pp. 1261-1266, July 2006.

[10] M. A. Salam, K. M. Jen and M.A. Khan, "Measurement and simulation of grounding resistance with two and four mesh grids," Power Electronics and Drive Systems (PEDS), 2011 IEEE Ninth International Conference on , pp.208-213, Dec. 2011.

[11] I. A. Letia, O. Matei, “Hybrid Neural Approach in Manura Robot Localization”, Advances in Electrical and Computer Engineering, vol. 5, no. 12, pp. 5-9, 2005.

[12] L. Neamt, O. Chiver, C. Barz, C. Cristinel, Z. Erdei, “Considerations about power system grounding for different soil structure”, Proceedings of the International Conference and Exposition on Electrical and Power Engineering, Iasi, Romania, 16-18 October 2014,.

[13] J. Schürmann, Pattern classification: a unified view of statistical and neural approaches. Wiley, New York, 1996.

[14] E. B. Baum, and D. Haussler. "What size net gives valid generalization?." Neural computation vol. 1, no.1, pp. 151-160, 1989.

[15] A. Krogh, J. Vedelsby. "Neural network ensembles, cross validation, and active learning." Advances in neural information processing systems, pp. 231-238, 1995.

[16] R. Setiono, "Feedforward neural network construction using cross validation." Neural Computation vol. 13, no. 12, pp. 2865-2877, 2001.

[17] Y. Ikegaya, A. Gloster, et al. "Synfire chains and cortical songs: temporal modules of cortical activity." Science vol. 304 no. 5670 pp. 559-564, 2004.

[18] L. C. Principe, N. R. Euliano, and W. C. Lefebvre. Neural and adaptive systems: fundamentals through simulations, Wiley, 2000.

[19] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian learning." Proceedings of the 1997 international joint conference on neural networks. vol. 3. Piscataway: IEEE, 1997.

[20] C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.

[21] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” Proceedings of the IEEE International Conference on Neural Networks (ICNN), pp. 586-591, San Francisco, 1993.

A Framework for an Effective Information Security Awareness Program in Healthcare

A Case Study of Computer Game in Hospital Universiti Kebangsaan Malaysia

Arash Ghazvini
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia,
43600 UKM, Bangi Selangor, Malaysia

Zarina Shukur
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia,
43600 UKM, Bangi Selangor, Malaysia

***Abstract*—Electronic Health Record (EHR) is a valuable asset of every healthcare and it needs to be protected. Human errors are recognized as the major information security threats to EHR systems. Employees who interact with EHR systems should be trained about the risks and hazards related to information security. However, there are limited studies regarding the effectiveness of training programs. The aim of this paper is to propose a framework that provides guidelines for healthcare organizations to select an effective information security training delivery method. In addition, this paper proposes a guideline to develop information security content for awareness training programs. Lastly, this study attempts to implement the proposed framework in a selected healthcare for evaluation. Hence, a serious game is developed as a training method to deliver information security content for the selected healthcare. An effective training program raises employees' awareness toward information security with a long-term impact. It helps to gradually change employees' behavior over time by reducing their negligence towards secure utilization of healthcare EHR systems.**

***Keywords*—awareness Training Program; Information Security; Content Development; Electronic Health Record; Human Error; Serious Game**

I. Introduction

There is a wide range of training delivery methods for information security awareness programs. However, research is scant regarding the effectiveness of these methods [1]. The literature shows that many information security awareness training programs have failed to produce long-term impacts on employees' behavior [1][2][6]. The success of any information security awareness program heavily relies on how the content is communicated to its audience [7]. Therefore, it is vital to select the most suitable information security training delivery method [1].

Hence, there is a need to develop a guideline for healthcare organizations to select the most suitable training delivery method for implementation of information security awareness program. It is important to develop a good training delivery method for information security awareness program that is accepted by employees to promote their participation. Furthermore, it is necessary to ensure that information security content is created through a valid process and effectively communicated with employee during awareness training program. Constructing a good information security awareness training program highly relies on the selected delivery method, design of the selected delivery method and training content.

Training Method Selection (TMS) framework is proposed as guideline for healthcare organizations to select an effective training delivery method for information security awareness. The key attributes of effective information security awareness training program are identified to be used in the proposed framework. In addition, the involvement of healthcare decision makers is important to use their insights in order to develop the framework. Hospital University Kebangsaan Malaysia (HUKM) is the selected healthcare for the purpose of this study. A serious game called InfoSecure is developed to test the effectiveness of the TMS framework. In the process of content development, Common mistakes made by healthcare employees were collected through questionnaire and wrong answers were used in development of training content. According to the findings, this study confirms that the selected training delivery method for HUKM produces desirable outcomes and accepted by both organization and employees.

II. Literature Study

The aim of this section is to find the gap and identify issues by understanding the background and previously conducted researches that may lead to potential solutions. There are three fundamental questions to be answered in this section as follow:

The first fundamental question is" how to select effective training method for information security awareness training program?" Health organization faces many challenges when it comes to selecting and effective information security awareness training program. The literature shows most of these programs failed in the past due to different reasons. A guideline is needed to assist organizations in order to select a training delivery method for information security. The Training Method Selection (TMS) is proposed as a solution to assist healthcare organization to select the most suitable information security awareness training delivery method. The TMS framework is influenced by Morrison et al. (2004), Manke and Winkler (2012), and Kissack and Callahan (2010) [17][15][13]. TMS framework is validated by expert panel approach and tested at Hospital University Kebangsaan Malaysia (HUKM) as a selected healthcare for this study.

Based on inputs from HUKM decision makers, computer game is selected as the most suitable information security awareness training delivery method for the organization. However, TMS framework may result different if utilized by other healthcare organization.

To answer the second fundamental question, “How to construct a good training delivery method for information security awareness training program?”. A well-designed layout, guideline, conceptual model is needed to construct a successful training program. This model can be formed based on models developed in previous studies. Previous studies recognized serious game as an appropriate solution for information security awareness training program [18][16][6][20]. Although, developing a successful serious game requires a review of adequate guidelines that identified all characteristics to be incorporated in such games. The developed serious game for HUKM is called InfoSecure. The InfoSecure conceptual model is influenced by Yusoff (2010) [24]. InfoSecure is validated by expert panel and pilot test confirms the effectiveness of the game before implementation at HUKM. The main objective of this awareness training program is to raise HUKM employees’ knowledge towards information security.

The third fundamental question is “How to develop an information security content for awareness training program?” A guideline is needed to assist organizations to create training content from the sources such as internal policy documents and international standards to be used in the selected training delivery method. Based on previous studies, an information security training content development guideline is proposed to help healthcare organization create information security content to be used as training material. As HUKM information security policy is outdated at the time this research taking place, the policy document is augmented based on information security international standards to be used as reference to create the content. Information security content of this study is validated by expert panel before added into the InfoSecure game. Format of content depends on the selected awareness training method. The result of content development for this study to be used in InfoSecure is 40 multiple choice questions and answers. All forty questions where given to selected number of HUKM employees as open ended questioner in order to collect their common mistakes as collection of wrong answer to be used in InfoSecure.

III. Research Aim

The Effectiveness of an information security awareness program has often been ignored by organizations. Electronic Health records (EHR) are the most valuable assets of every healthcare and it needs to be protected. Human errors are recognized as the major threats to electronic health record systems. Employees who interact with the systems must be trained to understand the risks associated with information security in EHR systems.

There is a wide range of information security awareness techniques. However, research is scant regarding effective information security awareness delivery methods. Although information security training programs can minimize the risks of employees’ mistake, the literature shows that many awareness training programs are not effective in raising employees’ awareness toward information security. The failure is due to several distinct problems such as lack of employees’ willingness to participate. Hence, the main objective of this paper is to provide guidelines for healthcare organizations to implement a successful awareness training program that raises employees’ awareness toward information security with a long-term impact.

Training delivery method is the key in designing an effective awareness program for information security. Hence, as the first objective, this study proposes a training method selection (TMS) framework to select an effective training method for information security awareness program. It guides organizations to select the most suitable training delivery method that fulfils organization training needs while promoting employees’ engagement and increasing their interest. To meet this objective, semi-structured interviews were conducted at the selected healthcare to obtain insights from decision makers to develop the framework. The TMS framework is implemented in Hospital University Kebangsaan Malaysia as the selected healthcare. Previously conducted awareness training programs at HUKM did not produce desired outcome. Based on the TMS framework, the healthcare decision makers selected computer game method for awareness training program. HUKM requires a fun, creative training programs that covers all employees and can be conducted more frequently. Moreover, it was required that the training program should be organized in a friendly and informal manner and lasts for approximately 30 minutes. In addition, the result shows that the most common information security incident occurring in the organization include 1) phishing, 2) web using, 3) email and spam, 4) malicious code, 5) password protection, 6) privacy and confidentiality, 7) workstation and hacking, and 8) access control.

As the second objective, this study develops and implements a serious computer game for HUKM to deliver the training content. Serious games are a type of computer games designed for training purposes that bring education and entertainment together. A serious game consists of educational elements with pleasant interface. The developed serious game is called InfoSecure enhances previously developed games in a number of dimensions such as flexibility and fun. The findings indicate that serious games with combination of two genres, simulation and casual, produce satisfactory outcomes. Simulation characteristic of a game allows users to make mistakes and learn from those mistakes without worrying about the consequences of their actions as they would in the real life. Casual characteristic provides flexibility and fun required in serious games. Hence, a combination of the two genres together results in a better serious game.

As the third objective, this study proposes guidelines to develop information security content for awareness training program. Training content must be developed based on i) healthcare internal information security policy, ii) information security international standards, iii) common information security mistakes made by employees, iv) selected training delivery method, and iv) targeted audience profile. It is realized that HUKM internal policy document, in some parts, is not in line with international standards. Therefore, this study proposes policy augmentation for HUKM. Subsequently, training content is developed for HUKM based on the augmented policy document. The main objective of training content is to enforce HUKM information security policy document.

IV. Training Delivery Method Framework

According to Holton (1996), the main failure of training programs is training design [10]. Training delivery method has a direct influence on success of training program [17]. The key to enhance an effective training program is to select an effective training delivery method [17]. In many cases, awareness training seem less likely to enhance employees' performance and they fail to produce satisfactory outcomes [2]. Even though it is necessary to ensure that an information security awareness program covers appropriate topics, it is important to select suitable training delivery methods [1]. Similar to any other programs, the success of an information security awareness program heavily depends on the way awareness information is communicated [7]. There are various types of training delivery methods for information security awareness program adapted by organizations [1]. Even though studies have been carried out to examine the efficiency of training delivery methods, research is scant regarding the effectiveness of the training delivery methods [1]. This study intends to fill this important gap.

The study proposes a training method selection (TMS) framework that helps organizations to select the most effective training delivery method for information security awareness program. The Training Method Selection (TMS) framework is developed based on i) Kemp instructional design model, ii) literature study, and iii) interview at selected healthcare. Kemp' model emphasizes on the importance of training delivery method. In addition, previous studies have identified important attributes that affect the effectiveness of delivery training methods, as discussed in the literature. Moreover, semi-structured interviews are conducted in a selected healthcare to obtain insights from decision makers. Figure 1 present the TMS framework.

A. Common information security issues

The first step in designing a training instruction is to identify a problem. Therefore, the initial step to design an information security awareness training program is to identify common information security issues that frequently occur by employees in healthcare organizations. These issues can be found in literature.

B. Selected information security topics

Different organizations deal with different information security issues. Therefore, organizations need to identify specific issues to be addressed in awareness training program. Moreover, content sequencing is regarded important as it helps learners understand information and materials easily. Therefore, ordering information security awareness materials by security topics help employees understand the idea properly and effectively.

C. Training content

Training content is important to help learners understand information and materials easily. Training content must be developed based on i) organization internal information security policy, ii) information security international standards, iii) common information security mistakes made by employees, iv) selected training delivery method, and iv) targeted audience profile. Organization internal information security policy is an important item to cover in training content [25][26][12]. Addressing internal policy helps employees understand the importance of information security and learn how to prevent incidents from happening. Moreover, looking at employees' common mistake help to determine the level of security awareness training required for organization to avoid over-train or under-train employees [25]. It is important to ensure that employees understand the delivered content; otherwise they may involuntarily put corporate information at jeopardy. Therefore, getting feedback on training content and employees' level of understanding are the keys to confirm personnel comprehension of the content as well as corporate security policy.

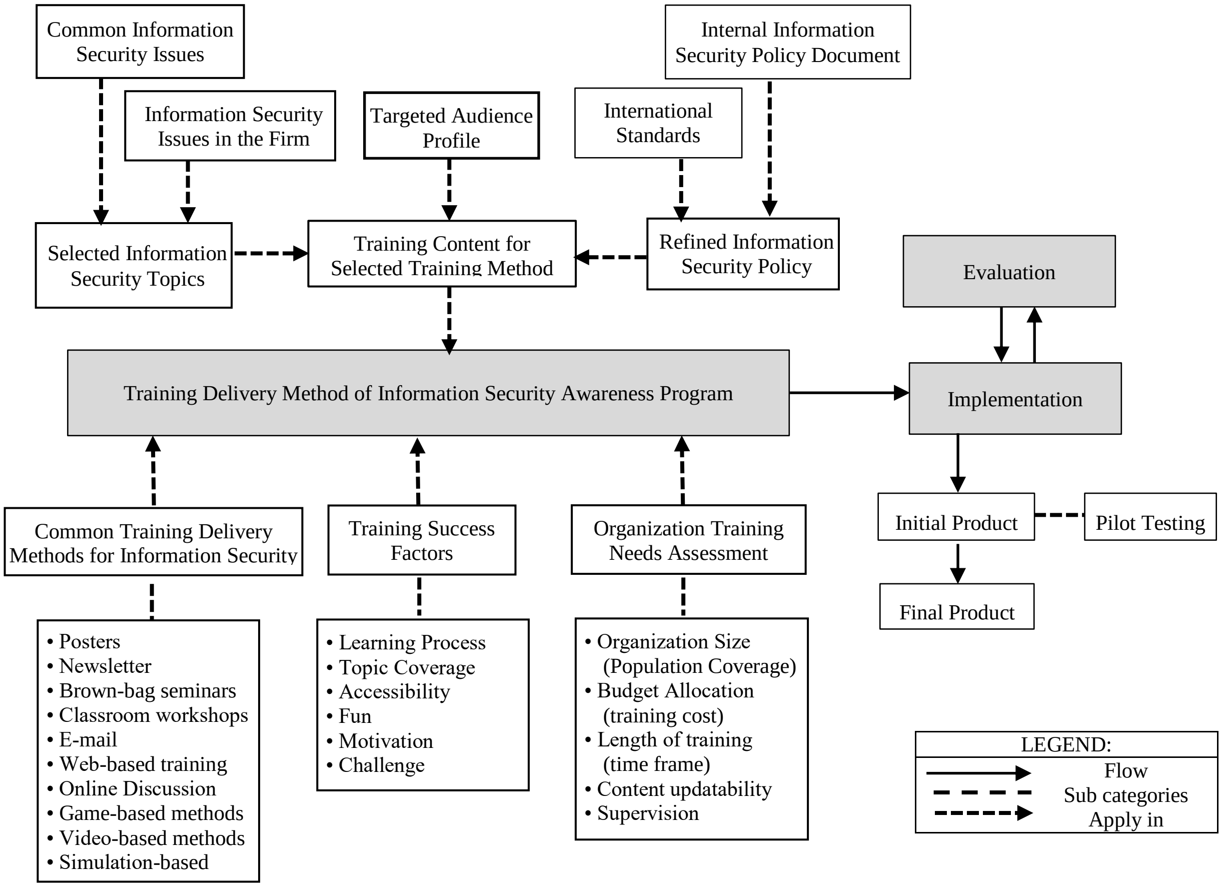

Common Information Security Issues
Information Security Issues in the Firm
Targeted Audience Profile
International Standards
Internal Information Security Policy Document
Selected Information Security Topics
Training Content for Selected Training Method
Refined Information Security Policy
Evaluation
Training Delivery Method of Information Security Awareness Program
Implementation
Common Training Delivery Methods for Information Security
Training Success Factors
Organization Training Needs Assessment
Initial Product
Pilot Testing
Final Product
• Posters
• Newsletter
• Brown-bag seminars
• Classroom workshops
• E-mail
• Web-based training
• Online Discussion
• Game-based methods
• Video-based methods
• Simulation-based
• Learning Process
• Topic Coverage
• Accessibility
• Fun
• Motivation
• Challenge
• Organization Size (Population Coverage)
• Budget Allocation (training cost)
• Length of training (time frame)
• Content updatability
• Supervision
LEGEND:
Flow
Sub categories
Apply in

Fig. 1. Training method selection (TMS) framework

D. *Refined information security policy*

The initial step in developing training content is to identify common information security issues in organization. Next step is to ensure that their internal information security policy document is up-to-date and in line with international standards. Therefore, they need to review and refine the existing internal policy based on the international standards if necessary. Training content development process will be elaborated in details in section X-XI.

E. *Targeted audience profile*

An information security content must be developed taking to account the targeted audience profile. If the massage is deigned too hard to understand, it will drive beginners away and if too easy will make professionals bored. Looking at targeted audience profile help to determine the level of security awareness training required for organization to avoid over-train or under-train employees.

F. *Common training delivery methods for information security*

A critical step in the instructional design process is to select the most appropriate training delivery method. The choice of raining delivery methods has significant impact on individual performance. There are a wide range of training delivery methods. However, reference [1] identified the most proper training delivery methods for information security. The list include i) paper-based (posters and newsletter), ii) instructor-led (brown-bag seminars and classroom workshops), iii) online (e-mail, web-based training, and online discussion), iv) game-based, v) video-based, vi) simulation-based. Narrowing down the list to only those applicable for information security, makes it easier for organization to recognize the most suitable method.

G. *Training success factors*

A well-developed instructional strategy motivates and attracts learners to training information. The training success factors include learning process, topic coverage, accessibility, fun, motivation, and challenge. The key to enhance successful awareness training program is to ensure that the program addresses employees' needs and preferences and it promotes employees' engagement to training activities.

H. *Organization training needs assessment*

The instructional objectives provide a map for designing the instruction and for developing the means to assess learner performance. Therefore, organization training need including population coverage, training cost, time frame, content updatability, and supervision. For instance, organizations may require post-training, hence, content updatability allows trainers to change or edit training content. It is important to ensure that developed instruction solve individual performance.

V. Training Delivery Method Map

To make the TMS framework easy to understand and use, a TMS map (Table 1) is developed that works as a check list for healthcare organizations. Decision makers can select an effective awareness training delivery method based on the TMS map. The left hand side column lists the most proper information security training delivery methods. The rest of the columns are classified into three categories; training success factors, organization training need, training content. Each category consists of several components. If a training method offers a component, it is indicated by **√** mark, if not by × mark.

Decision makers need to carefully select those elements from the TMS map that are most critical to their organization. For example, large population coverage might be very critical to training need of a large organization but less important for small organizations. The training delivery method that has **√** mark for those selected elements is the most suitable method for that organization.

TABLE I. Training Method Selection (TMS) Map

Training Delivery Methods		Training Success Factors						Organization Training Need					Feedback on Content Training
		Active Learning Process	Multiple Topic Coverage	Easily Accessible	Fun	Motivation	Challenge	Large Population Coverage	Low Cost	Flexible Time Frame	Content Updatability	Supervision	Question & Answer
Paper-based Methods	Posters	×	×	×	×	×	×	√	√	√	×	×	×
	Newsletter	×	×	×	×	×	×	√	√	√	×	×	×
Instructor-led methods	Brown-bag seminar	×	×	×	×	√	×	×	√	×	√	√	√
	Classroom workshop	×	√	×	×	×	×	×	√	×	√	√	√
Online methods	E-mail	×	×	√	×	×	×	√	√	√	×	×	×
	Web-based training	√ ×	√	√	×	×	√ ×	√	√	√	√	√ ×	√
	Online Discussion	√	√	√	√	√	√	×	√	×	√	√	√
Game-based method		√	√	√	√	√	√	√	×	√	√ ×	×	√
Video-based method		×	√	√	×	×	×	√	×	√	×	×	×
Simulation-based method		√	×	×	√	√	√	×	×	×	√	×	√

VI. TMS Framework Validation

An expert panel approach is used to evaluate the TMS framework. Panel of experts is the initial and critical step in establishing content validity. In this study, three professors reviewed the framework and semi-structured interview questions for the purpose of ensuring language, wording, layout, importance of the framework, relevancy of the framework to objectives, process of content creation, clarity of point, topic coverage of the framework and semi-structured interview questions. The expert panel provided the candidate with their inputs, correction and area of improvements.

The expert panel appraisal developed by the researcher is based on (a) literature, (b) requirement of the study, (c) guideline suggested by [14][22]. The panel was asked to provide their recommendations of amendments based on the provided questions. Finally, based on their discussion and comments, the framework and semi-structured interview questions are modified and further improved.

The evaluation confirms the followings: a selected training method based on TMS framework could be the most suitable training method for the selected organization; the selected training method based on TMS framework could promote employees' motivation and participation in training activities; information security content created based on TMS framework could effectively deliver the training content to employees; and the program could successfully enhance employees' knowledge towards information security and strengthens their understanding of organization's information security policy.

VII. TMS Framework Implementation

Hospital Universiti Kebangsaan Malaysia (HUKM) is selected as the case healthcare to implement and evaluate the TMS framework. HUKM has implemented number of awareness training programs that failed to produce satisfactory outcome. The case study approach is now widely used, and there is a growing confidence in its applicability to examine a new finding within real-life context [8][23].

Semi-structure interviews were conducted with key decision makers of the selected hospital. Multiple interviews were conducted with the heads of IT department to identify common issues and mistakes made by employees as well as discussing the training success factor and organization needs assessment. Delphi method is utilized to conduct interviews as it limits the range of respondents' feedbacks and helps coverage toward correct answers. The inputs obtained from the semi-structured interviews intend to evaluate the framework and also to identify which training method is best approach for HUKM. The questionnaire consists of clear and concise instructions divided into four parts.

Eventually, HUKM decision makers converged to the conclusion that computer game-based training is an effective awareness training program to raise employees' awareness toward information security in HUKM. Computer game-based training is perceived as an engaging approach to enhance employees' awareness toward information security. Computer game was selected as it fulfills the organizations' need and promotes employees' engagement.

VIII. Serious Game

Serious game refers to a game that is primarily designed for training purposes rather than pure entertainment. Computer games designed for training purposes should integrate educational content with multimedia while providing pleasant interface [19]. Reference [4] discussed the educational advantage of serious game. "The serious games application is intended to help professionals, as well as enabling users to enjoy themselves through straightforward, real interaction while learning how to cope in several real social situations".

The result of TMS framework revealed that computer game is the most suitable training delivery method for HUKM. Development process needs proper guidelines that comprise all characteristics that should be included in a serious game. Hence, it is important to review available serious game models. This study develops InfoSecure conceptual model based on the model proposed by Yusoff (2010) (Figure 2) as it is the most efficient and effective model for serious games [5]. As Yussof (2010) explains, **capability** is player's capability to be learned in the game. **Instructional content** refers to the subject matter that player is required to learn. Both capability and Instructional content are components of **intended learning outcomes**, which refer to the objective playing a serious game. **Game attributes** are game functions that support players learning and engagement. Game attributes and intended learning outcomes are components of **learning activity**. G**ame genre** refers to style and characteristics a game that specifies the type of environment for the set of activities to be played within the game world. G**ame mechanics** are the components that offer more enjoyment and engagement to a game.

IX. The InfoSecure Conceptual Model

Figure 3 is the developed conceptual model for InfoSecure game. Because this is a conceptual model, it can be used as a guideline that visually represents the arrangement of the InfoSecure game elements.

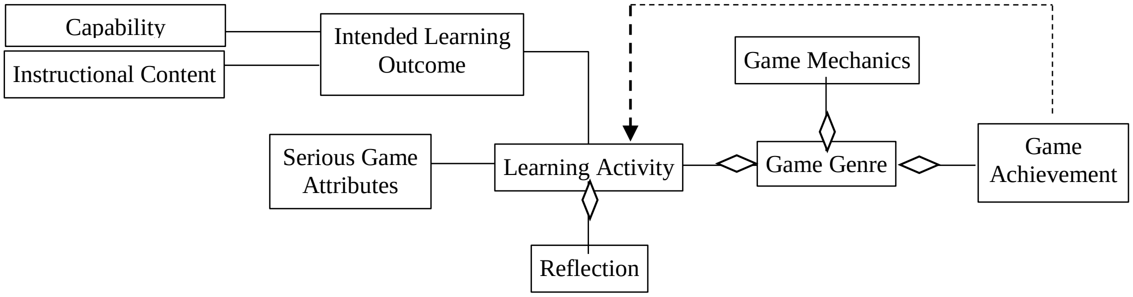


Fig. 2. Yusoff's (2010) serious game model

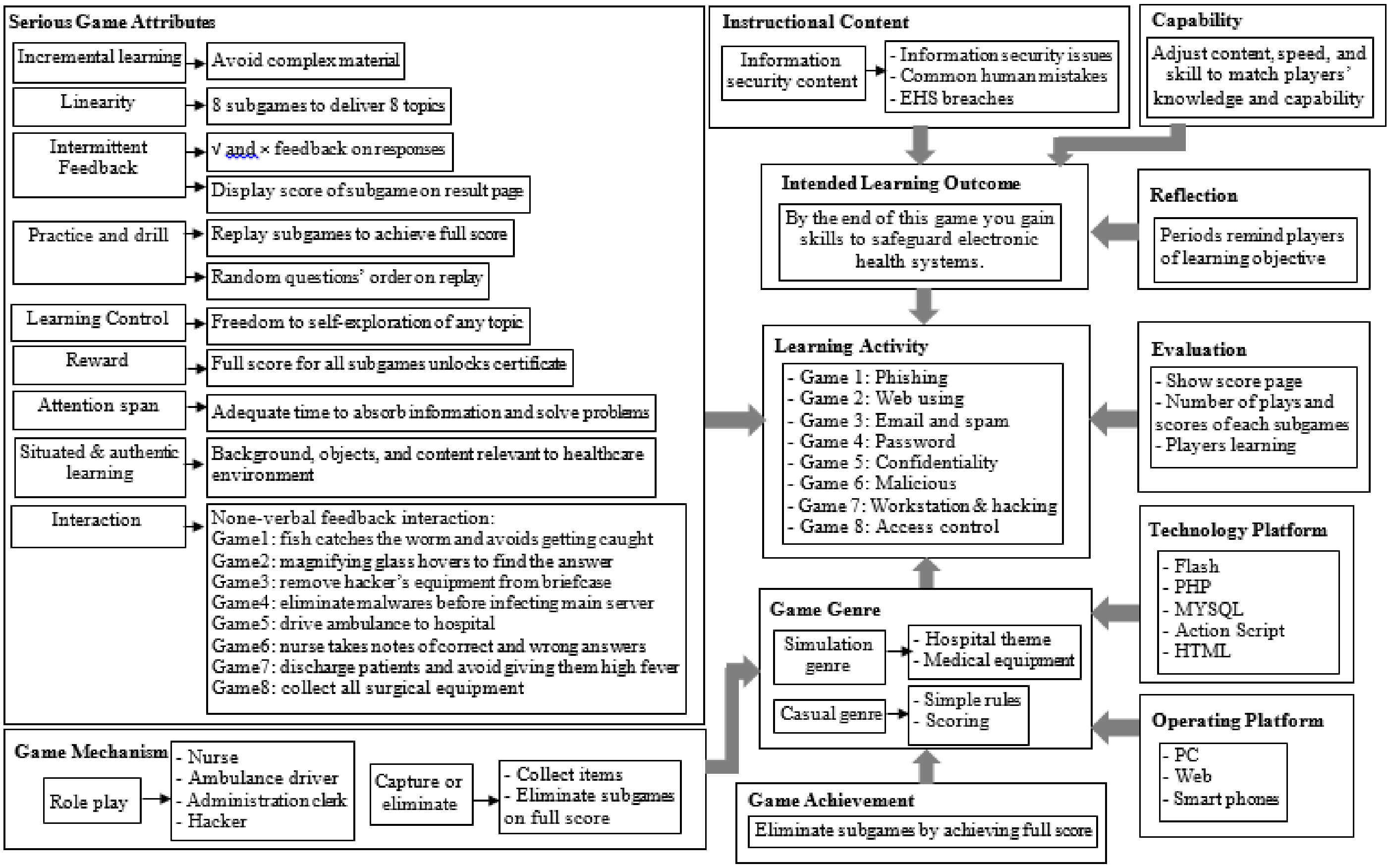

Serious Game Attributes
Incremental learning
Avoid complex material
Linearity
8 subgames to deliver 8 topics
Intermittent Feedback
√ and × feedback on responses
Display score of subgame on result page
Practice and drill
Replay subgames to achieve full score
Random questions' order on replay
Learning Control
Freedom to self-exploration of any topic
Reward
Full score for all subgames unlocks certificate
Attention span
Adequate time to absorb information and solve problems
Situated & authentic learning
Background, objects, and content relevant to healthcare environment
Interaction
None-verbal feedback interaction:
Game1 : fish catches the worm and avoids getting caught
Game2: magnifying glass hovers to find the answer
Game3: remove hacker's equipment from briefcase
Game4: eliminate malwares before infecting main server
Game5: drive ambulance to hospital
Game6: nurse takes notes of correct and wrong answers
Game7: discharge patients and avoid giving them high fever
Game8: collect all surgical equipment
Game Mechanism
Role play
- Nurse
- Ambulance driver
- Administration clerk
- Hacker
Capture or eliminate
- Collect items
- Eliminate subgames on full score
Instructional Content
Information security content
- Information security issues
- Common human mistakes
- EHS breaches
Capability
Adjust content, speed, and skill to match players' knowledge and capability
Intended Learning Outcome
By the end of this game you gain skills to safeguard electronic health systems.
Reflection
Periods remind players of learning objective
Learning Activity
- Game 1: Phishing
- Game 2: Web using
- Game 3: Email and spam
- Game 4: Password
- Game 5: Confidentiality
- Game 6: Malicious
- Game 7: Workstation & hacking
- Game 8: Access control
Evaluation
- Show score page
- Number of plays and scores of each subgames
- Players learning
Technology Platform
- Flash
- PHP
- MYSQL
- Action Script
- HTML
Game Genre
Simulation genre
- Hospital theme
- Medical equipment
Casual genre
- Simple rules
- Scoring
Operating Platform
- PC
- Web
- Smart phones
Game Achievement
Eliminate subgames by achieving full score

Fig. 3. InfoSecure conceptual model

X. Policy Augmentation

HUKM has not developed its own information security policy document, and thus, it follows Universiti Kebangsaan Malaysia's (UKM) information security policy. Even though HUKM is ISO certified, there are sections outdated or insufficient in details in the policy. Therefore, to develop appropriate training content, this study augments HUKM information policy document based on relevant international standards. The aim of augmentation is to encourage policy makers in HUKM to update current internal information security policy and to use the augmented document to create future content for post-trainings. To augment HUKM internal policy document, the researcher reviewed and refined the document using international standards including ISO 27002, SANS, and HIPAA.

For demonstration purposes this paper only shows policy augmentation process for "control of logical access" policy. As shown in Table 2, the column on the left is a list of topics selected from HUKM policy document and international standards. The other four columns are selected sources including HUKM, ISO, SANS, HIPAA. The first step is to gather a list of topics. The researchers reviewed HUKM policy as well as the other three sources to identify missing information in HUKM document. For instance, as table shows, HUKM does not have any policy regarding workstation use. The next step is to distinguish which of the topics are covered by each source, as indicated by **√** in the table. Subsequently, the strength and quality of policy statement provided by each source was carefully evaluated in comparison with HUKM's policy statements. Policy statements were extracted from sources and incorporated into HUKM policy document when necessary, as indicated by [**√**].

A. Control of Logical Access

The objective is to safeguard healthcare information assets including electronic health record from unauthorized access. Security facilities are required to prevent unauthorized access to health information systems. Logical access to health information systems should be only given to authorized individuals. Table 3 proposes policy augmentation for control of logical access.

TABLE II. POLICY AUGMENTATION

Topics	HUKM	ISO 27002 2005	SANS	HIPAA
Server Security				
Physical Security Control	[√]	√		
Control of Database	[√]	√		
Control of Logical Access	[√]	√		
User Identification	[√]	[√]	√	√
User Authentication	[√]	[√]	[√]	√
Information Back-up	[√]	[√]		
Maintenance	[√]	[√]		
Workstation Use			√	[√]

TABLE III. CONTROL OF LOGICAL ACCESS POLICY AUGMENTATION

HUKM Policy	Augmentation Source: ISO 27002 2005; Sec 11.5.3	Augmentation Source: SANS; Password Policy
User Identification 1. System users are individual or group of users that share the same user account and is responsible for the security of the system used. HUKM identify illegal users through the following steps: a. Given one (1) unique ID to all individual user; b. Store and maintain all user ID responsible for each activity; c. Make sure there is auditing facility to check all user activity; d. Make sure all created user ID is based on application; and e. Changes of user ID for application software must get permission from that Application Systems' Secretariat. 2. HUKM identify inactive user ID are not misused through the following steps: a. Suspend all unused ID facilities for 60 days and delete the ID after the 60 days period, and b. Delete all facilities for users that have moved department or retired; *User Authentication* The system should be able to provide the following facilities: 1. The password entered in the form of not visible; 2. The length of password must be at least eight (8) characters long with combination of characters, numbers or other symbols; 3. The password is encrypted during submission; 4. Password file is kept apart from the data for main application system; and Access attempt is limited to five (5) times only. The user ID must be suspended after five (5) consecutive times of trial	*Password Management System* A password management system should: 1. Enforce the use of individual user IDs and passwords to maintain accountability; 2. Allow users to select and change their own passwords and include a confirmation 3. procedure to allow for input errors; 4. Enforce a choice of quality passwords; 5. Enforce password changes; 6. Force users to change temporary passwords at the first log-on; 7. Maintain a record of previous user passwords and prevent re-use; 8. Not display passwords on the screen when being entered; 9. Store password files separately from application system data; 10. Store and transmit passwords in protected form (e.g. encrypted or hashed) *Password Use* 1. Keep passwords confidential 2. Avoid keeping a record 3. Change passwords whenever there is any indication of possible system or password compromise 4. Select quality passwords with sufficient minimum length 5. Not vulnerable to dictionary attacks 6. Free of consecutive identical, all-numeric or all-alphabetic characters. 7. Change passwords at regular intervals and avoid re-using or cycling old passwords 8. Change temporary passwords at the first log-on 9. Not include passwords in any automated log-on process 10. Not share individual user passwords Not use the same password for business and non-business purposes	*Password Construction Guidelines* All users at HUKM should be aware of how to select strong passwords. Strong passwords have the following characteristics: 1. Contain at least three of the following character classes: a. Lower case characters b. Upper case characters c. Numbers d. Punctuation e. Special characters f. Contain at least fifteen alphanumeric characters. C1. Weak passwords have the following characteristics: a. The password contains less than fifteen characters b. The password is a word found in a dictionary (English or foreign) The password is a common usage word such as: names of family, pets, friends; the words "HUKM, PPUKM"; birthdays and other personal information such as addresses and phone numbers; word or number patterns like aaabbb, qwerty, zyxwvuts, 123321, etc.

XI. TRAINING CONTENT FOR HUKM

This section explains the process to create training content for HUKM from the augmented policy document. In what follows, information security questions and answers are described.

A. The Questions

As discussed earlier, the TMS framework was implemented at HUKM and it is found that computer game is the most suitable training delivery method for this healthcare. The purpose of this section, is to develop training content in the form of information security questions. A total of 40 questions are created based on HUKM augmented policy document.

B. The Wrong Answers

This study conducted a survey among healthcare employees to collect employees own wrong answers to information security questions. The wrong answers are, used to design the training content. This approach is helpful i) to understands the knowledge level of employees about security topics, iii) to address employees' real problem in understanding information security topics, and ii) to mislead employees and to evaluate their real understanding of subject matters. For this purpose, information security questions are prepared in form of open-ended structure and they are distributed among employees. For example, many employees responded that the minimum length of strong password is four characters whereas the right answer is eight characters.

C. The Correct Answers

The correct answers, on the other hand, are extracted form HUKM augmented policy document, because the objective of training content is to enforce HUKM policy document. For instance, on user authentication topic, the minimum length of strong password is eight characters as stated in HUKM policy document. However, ISO suggests that a strong password must contain at least fifteen characters. Although HUKM is ISO certified, the correct answer to choose should be minimum of eight characters. However, only few sections of HUKM policy document have insufficient information on some of the selected topics. Therefore, some of the questions and correct answers are taken from international standards and verified by the healthcare. Since HUKM is ISO certified, ISO 27002 is prior to other international standards. Table 4 presents the questions and answers created for password protection.

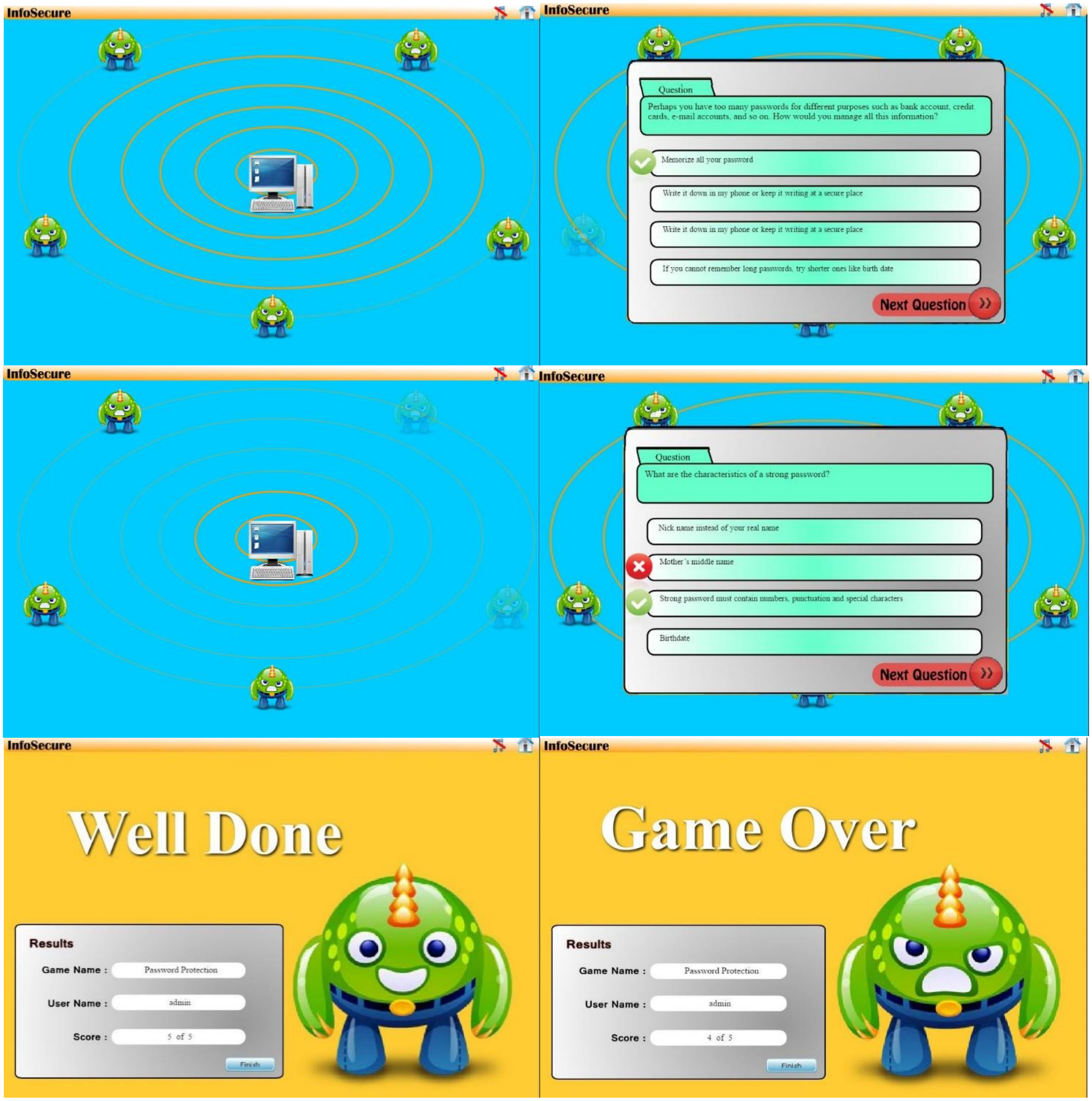
TABLE IV. QUESTIONS AND ANSWERS FOR PASSWORD PROTECTION

Question	Wrong Answer	Correct Answer
Q1. What do you think the minimum length of a strong password should be (e.g. 5 Characters)? S1. HUKM Security Policy 2.3.2 (2)	W1. Four W2. Six W3. Ten	C1. At least eight alphanumeric characters.
Q2. What are the characteristics of a strong password? S1. HUKM Security Policy 2.3.2 (2)	W1. Nick name instead of your real name W2. Mother's middle name W3. Birth date	C1. A strong password contains at least three of the five following character classes: - Lower case characters - Upper case characters - Numbers - Punctuation - Special characters (e.g. @#$%^&*()_+|~-=\`{}[]:";'<>/ etc) - Contain at least eight alphanumeric characters
Q3. What are the characteristics of a weak password? S1. HUKM Security Policy 2.3.2 (2)) S2. SANS; Password Policy (2)	W1. Contains only alphabet and numbers W2. Alphanumerical password W3. Contain at least eight alphanumeric characters password	C1. Weak passwords have the following characteristics: - The password contains less than fifteen characters - The password is a word found in a dictionary (English or foreign) - The password is a common usage word such as: names of family, pets, friends; the words "HUKM, PPUKM"; birthdays and other personal information such as addresses and phone numbers; word or number patterns like aaabbb, qwerty, zyxwvuts, 123321, etc.
Q4. Your colleague calls you from home to ask your staff ID and password. What should you do? S1. ISO 27002 2005; Sec 11.5.3 Password Use (1, 2 &10)	W1. Ask for the reason before revealing the password W2. Only reveal the password in case of an emergency and change it afterwards. It is okay if you know the person.	C1. All users should be advised to not share individual user passwords. Do not share HUKM passwords with anyone, including administrative assistants or secretaries.
Q5. Perhaps you have too many passwords for different purposes such as bank account, credit cards, e-mail accounts, and so on. How would you manage all this information? S1. ISO 27002 2005; Sec 11.5.3 (Password Use (1, 2 &10)	W1. Use same password and remember it. W2. If you cannot remember long passwords try shorter ones like birth date. W3. Write it down in my phone or keep it writing at a secure place	C1. Memorize all your password C2. Avoid keeping a record (e.g. paper, software file or hand-held device) of passwords, unless this can be stored securely. C2. Passwords should never be written down or stored on-line without encryption.

Note: Q stands for question; S stands for source; W stands for wrong answer; C stands for correct answer

XII. THE INFOSECURE GAME FOR PASSWORD PROTECTION

This paper develops a serious game called InfoSecure as a training tool to deliver the developed information security content. The InfoSecure game consists of 8 subgames each covering an individual topic. For demonstration, figure 4 shows screenshots of an InfoSecure sub-game that covers password protection. The story of the game is to remove all viruses before reaching the main server by answering all questions correctly. InfoSecure is a dynamic game and not static. That is, instructors are able to change and customize the training content as well as the graphics. Instructors with administrative privilege are able to determine the number of questions from a range of one to ten.



Perhaps you have too many passwords for different purposes such as bank account, credit cards, e-mail accounts, and so on. How would you manage all this information?

✓ Memorize all your password

Write it down in my phone or keep it writing at a secure place

Write it down in my phone or keep it writing at a secure place

If you cannot remember long passwords, try shorter ones like birth date

What are the characteristics of a strong password?

Nick name instead of your real name

✗ Mother's middle name

✓ Strong password must contain numbers, punctuation and special characters

Birthdate

Results

Game Name : Password Protection

User Name : admin

Score : 5 of 5

Results

Game Name : Password Protection

User Name : admin

Score : 4 of 5

Fig. 4. Screenshots of password protection game

There are two icons on top right corner: mute/unmute and home button. The home button redirects user to the main page to whether replay the same game or play a different game. Below the top banner, there is a green color dice on upper right side of the screen. Every time user clicks to roll the dice a new question displays and the total number of questions are set to the number of questions determined by trainer. Once a question is answered correctly, a green color ✓ mark appears indicating that correct answer is chosen, and the virus icon fades away. If user selects a wrong answer, a red color × mark appears indicating that user selected a wrong answer, and at the same time, the correct answer will be shown by a green color ✓ mark. Once an answer is select, whether correct or wrong, user can click on next question button and proceed with rest of the questions. User is required to answer all questions correctly otherwise he/she has to replay the game until all questions are answered correctly and the game is marked as completed on the home page. To prevent users from memorizing the patterns of correct answer, the order of questions randomly changes whenever a game starts. Once all questions are answered, a result page appears that displays game topic, username and score.

XIII. INFOSECURE PILOT TEST

A pilot test was conducted as a preliminary trial for the InfoSecure game. Pilot test helps researchers to identify possible flaws or weaknesses in a product. Ten participants for the pilot study, who volunteered to play the InfoSecure game, were randomly selected and divided into two groups. Five multimedia students from UKM made up the first group, while another twenty HUKM employees made up the second group.

Users who play a subgame for the first time would answer information security questions based on their initial knowledge and understanding, which may be incorrect. A player who answers two questions correctly and scores 40% when playing a subgame for the first time would know that he has answered three questions incorrectly. The player then has to repeat the subgame. In the second attempt, the player would be more cautious in answering the questions, having known that some of the previously selected answers had been incorrect.

With the assumption that the player manages to answer four questions correctly and scores 80%, he still has one question to answer, and thus has to play the subgame all over again. The subgame cannot be deactivated and marked as completed until all five questions have been correctly answered and a 100% score is attained. Hence, the subgame has to be replayed until all of the questions are correctly answered. Once the 100% score is attained, the subgame will deactivate and will be marked as completed. The order of the questions changes every time the subgame is replayed so that a player is prevented from memorizing the sequence and pattern of the correct answers. This is achieved by shuffling the questions so that they are displayed in a random order every time the subgame restarts.

Employees gain two benefits by playing the InfoSecure game; it helps them to gain knowledge on information security, and to replace the incorrect information they might initially have in their minds prior to playing the game. It also helps them to understand the importance of thinking carefully when dealing with the electronic health systems. In the game, users are allowed to make mistakes and learn from them without having to worry about the consequences of their actions as they would in real life.

A player's (employee) score for each subgame, which includes his very first attempt until his last attempt in attaining the 100% score, would be recorded in a database. User progress is displayed in the player's profile, which can be viewed by both the player and the instructors. This helps the instructors to keep track of the employees' performance which demonstrates their learning curve. The aspects of player performance include the frequency of a subgame being played, the scores attained, the most difficult information security topics, and also the employees' strengths and weaknesses. The ability to evaluate the recorded information helps managers in monitoring their employees' performance and in taking the necessary actions. Players who scored 100% in all the subgames will be awarded with a certificate of accomplishment, which can be printed once the game is completed.

Nevertheless, obtaining 100% score is not the ultimate goal since it is crucial for the employees to fully understand the topics and to integrate them in their daily activities. Therefore, the game must be played frequently, as determined by the hospital management. In line with the aim of keeping the gameplay more interesting, and to avoid the reuse of static games and to maintain the players' motivation in taking part in the game, InfoSecure is developed to be dynamic by allowing IT managers to change and customize the training content as well as the graphics.

Getting the feedback on gameplay experience is the key objective in asking computer science students to participate in the game. It is not surprising that even during the first play of the game, the computer science students have managed to perform well by answering most of the questions correctly. HUKM employees on the contrary, had to play a subgame a few times before scoring 100%. Table 5 below shows sample of five employees' records for subgame covering password protection topic. Employee number 3 for example, played the phishing subgame four times. For the first play, he managed to score only 20% by getting one correct answer. For the second and third plays, the score had increased to 40% and 80% respectively. During the fourth play, the player managed to select the correct answers for all the questions and thus scored 100%.

TABLE V. EMPLOYEES' RECORD OF PLAYING INFOSECURE

Subgame	Employee #1	Employee #2	Employee #3	Employee #4	Employee #5
Phishing	1st play: 60% 2nd play: 100%	1st play: 80% 2nd play: 100%	1st play: 20% 2nd play: 40% 3rd play: 80% 4th play: 100%	1st play: 80% 2nd play: 80% 3rd play: 100%	1st play: 40% 2nd play: 80% 3rd play: 100%

The record shows that privacy and confidentiality, and workstation and hacking are the most challenging topics compared to other topics. The score of employee number three were both 0% when he played the above two games for the first time since none of his answers were correct. His score was also 0% when he first played the subgame on access control. Employees' total plays and their lowest and highest scores on their first attempt are shown in Table 6. The two subgames of privacy and confidentiality, and workstation and hacking were replayed more than the other subgames, each for a total of 18 times in order for the players to obtain a score of 100%. The lowest first play score goes to privacy and confidentiality (0%), workstation and hacking (0%), and access control (0%). The highest first play score goes to Phishing (80%), email and spam (80%), and access control (80%).

TABLE VI. LOWEST AND HIGHEST SCORES ON FIRST ATTEMPTS

Subgame	Total Play	Lowest Score	Highest Score
Phishing	14	20%	80%
Web using	16	20%	60%
Email and spam	15	40%	80%
Malicious code	14	20%	60%
Password protection	13	40%	60%
Privacy and confidentiality	18	0%	40%
Workstation and hacking	18	0%	40%
Access control	14	0%	80%

XIV. CONCLUSION

This research is a noteworthy attempt to address the issues concerning the effectiveness of information security awareness training programs. The research findings revealed the importance of training delivery method and training content in designing an effective information security awareness training program. Moreover, it is found that previously designed training programs failed because they were neither supported by organizations' needs nor accepted by employees. The findings in this study show that an effective information security awareness training program should be designed based on organizations' training needs while promoting employees' engagement and increase their interest.

Therefore, it is vital to give considerable attention to develop training delivery method and training content. Hence, this study proposed a training method selection (TMS) framework that helps organization to select an effective training delivery method for information security awareness program. The framework is based on the key attributes of effective training delivery method include training success factors and organization training needs. The selected training method based on the TMS framework is both supported by organization and accepted by employees.

By using the augmented information security document as the training content, and the computer game as the training tool, policy content was effectively delivered to employees to enhance their awareness. The result of this study reflects in enhancing employees' awareness toward the augmented information security policy in HUKM. An interactive computer game-based awareness training program gradually reduces employees' negligence and promotes the secure utilization of EHR system in HUKM to protect electronic health records.

Measuring employees' level of understanding of information security before and after the awareness training program indicates that the implemented program provides desired outcome. Employees have acquired better understanding of information security and they can manage to handle security matters in a way that limits damage and reduces recovery time and costs. The training program must be repeated frequently to keep employees updated and to change their habits over time. The success of the training program at HUKM shows that TMS framework is effective and it can be used as a guideline to select an effective training delivery method. Nevertheless, the TMS framework can be used by any healthcare to select, design and implement a successful information security awareness training program.

XV. RECOMMENDATION FOR FUTURE STUDIES

There is a wide range of information security awareness techniques. However, research is scant regarding effective information security awareness delivery methods. It is necessary for counselors, educators, and professionals to consider the findings obtained from the current study to further enhance awareness training programs. The findings can be used as a resource material for researchers, scientists, and university authorities who wish to conduct research in the same field. Therefore, the findings will help as supplementary evidence to obtain new results.

This study investigated a wide range of concepts to enhance effectiveness of information security awareness training program. Even though the findings of this study contribute to the field of information security, a few recommendations have yet to be provided for future research. Considering the depth and complexity of the topic there is room to explore and investigate more. Future researchers are recommended to study more elements affecting effectiveness of awareness training programs. Moreover, the findings of this study are limited to healthcare sector and are not generalizable to all organizations, Future studies are recommended to yield more representative active results.

REFERENCE

[1] Abawajy, J. 2012. User preference of cyber security awareness delivery methods. *Behavior & Information Technology* 33(3): 237–248.

[2] Annetta L.A. 2010. The "I's" Have It: A Framework for Educational Game Design. *Review of General Psychology* 14(2): 105-112.

[3] Apperley, T.H. 2006. Genre and Game Studies: toward a Critical Approach to Video Game Genres. *Simulation & Gaming*, 37(1).

[4] Bartolome, N.A., Zorrilla, A.M., Zapirain, B.G. 2011. Can game-based therapies be trusted? Is game-based education effective? A systematic review of the serious games for health and education. *The 16th International Conference on Computer Games*. University of Deusto.Avda. Universidades, Spain, 275-282.

[5] Buendía-García, F., García-Martínez, S., Navarrete-Ibañez, E.M. & Cervelló-Donderis, M.G. 2013. Designing serious games for getting transferable skills in training settings. *Interaction Design and Architecture(s) Journal* (19): 47-62.

[6] Cone, B.D., Irvine, C.E., Thompson, M.F., Nguyen, T.D. 2007. A Video game for cyber security training and awareness. *Computers & Security* 26: 63-72.

[7] Gardner, B., & Thomas, V. 2014. Building an information security awareness program: defending against social engineering and technical threats. *Elsevier*.

[8] Hartley, Jean. (2004). Case study research. In Catherine Cassell & Gillian Symon (Eds.), Essential guide to qualitative methods in organizational research, 323-333. London: Sage.

[9] HIPAA (The Health Insurance Portability and Accountability). 2014. www.hhs.gov/hipaa [10 January 2014].

[10] Holton, E. F. 1996. The flawed four level evaluation model. *Human resource development quarterly 7*(1): 5-21.

[11] ISO (International Organization for Standrdization) 27002. 2005. Standards. http://www.iso.org/iso/home/standards [23 February 2014]

[12] Johnson, E. C. 2006. Security awareness: switch to a better programme. *Network Security* 2: 15-18.

[13] Kissack H. C., Callahan J. L. 2010. The Reciprocal influence of organizational culture and training and development programs: building the case for a culture analysis within program planning. *Journal of European Industrial Training*, 34(4): 265-380.

[14] Iarossi, G. (2006). The power of survey design: A user's guide for managing surveys, interpreting results, and influencing respondents: World Bank Publications.

[15] Manke, S., Winker, I. 2012. The habits of highly effective security awareness program: A cross-computer comparison. *Internet Security Advisors Group*.

[16] Monk, T., Niekerk, J. & Solms R. 2010. Concealing the medicine: information security education through game play. *Institute for ICT Advancement, Nelson Mandela Metropolitan University*.

[17] Morrison, G. R., Ross, S. M., Kemp, J. E., & Kalman, H. (2004). *Designing Effective Instruction*: John Wiley & Sons.

[18] Nagarajan, A., Allbeck, J.M. & Sood, A. 2012. Exploring game design for cybersecurity training. *Proceedings of the 2012 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, Bangkok, Thailand*.

[19] Omar, H.M. & Jaafar, A. 2011. Usability of educational computer game (Usa_ECG): applying analytic hierarchy process. *International Visual Informatics Conference*,147-156.

[20] Prensky, M. 2001. True believers: digital game-based learning in the military. *From Digital Game-Based Learning, McGraw-Hill*, 2001.

[21] SANS (The System Administration, Networking, and Security, https://www.sans.org [22 May 2014].

[22] TrainingCheck. 2015. How can I pilot test the evaluation design and settings? http://www.trainingcheck.com/help-centre-2/faqs/evaluation-design-and-management/how-can-i-pilot-test-the-evaluation-design-and-settings/ [2 December 2015].

[23] Yin, R. K. (2003). Case study research: Design and methods. Sage Publications, Inc, 5, 11.

[24] Yusoff, A. 2010. A Conceptual Framework for Serious Games and its Validation. Thesis for the degree of Doctor of Philosophy. School Of Electronics and Computer Science, University 0f Southampton, United Kingdom.

[25] Security Standard Council. 2014. Information Supplement: Best Practices for Implementing a Security Awareness Program.

[26] Yanus R., &Shin. N. 2007. Critical success factors for mapping an information security awareness program. Proceedings of the Sixth Annual ISOneWorld Conference, Las Vegas, Nevada.

GSM based Android Application: Appliances Automation and Security Control System using Arduino

Kainat Fareed Memon, Javed Ahmed Mahar, Hidayatullah Shaikh, Hafiz Ahmed Ali, Farhan Ali Surahio
Department of Computer Science, Shah Abdul Latif University, Khairpur Mir's, Sindh, Pakistan

Abstract—**Now-a-days, automation is playing significant role using android phone in human life, particularly, handicapped and senior citizens. Appliances automation allows users to control different appliances such as light, fan, fridge and AC. It also provides security system like door controlling, temperature & fire detection and water shower. Furthermore, security cameras are used to control and monitored by the users to observe activity around a house. It has been observed that the internet services in interior Sindh are not as much better as required. Hence, GSM SIM900A based android application is developed named Appliances Automation & Security Control System using Arduino. The developed system is decomposed into two separate entities: (1) hardware is designed and developed using Arduino (MEGA 2560) with other required electronics components which is programmed using embedded C language, (2) an Android app which provides freedom to user to control and access the electronic appliances and the security system without internet. The developed application is tested in Karachi, Sukkur and Khairpur with ZONG, Moblink, Telenor and Ufone. The acceptable results are achieved at Karachi and Sukkur but suitable results are not calculated at Khairpur in terms of delay due to the frequency of selected GSM Module.**

Keywords—android application; gsm module; security system; Arduino

I. Introduction

The basic idea of home automation is observed since 1970s but expectations of the peoples are continuously and constantly increase due to the advancement of the technology and internet services. During the literature review, it has been observed that different researchers proposed architectures for various efficient and convenient home automation systems. Even the technology is entirely changed but the function and importance of home automation systems are same as previous [1] [2].

Recently, life is increasingly tight with the rapid growth in communications and information technology. The electronic and electrical environment with respect to automation of the household activities including centralized control of appliances, and other systems, to provide improved convenience, comfort, energy efficiency and security [3]. Appliances automation for the elderly and disabled can provide increased quality of life for persons who might otherwise require caregivers or institutional care. It can also provide remotely accessible environment in which each appliance can be remotely accessed and controlled using software as an interface, which includes an Android application [4].

The intellectual societies bring information where safe, economic, comfortable and convenient life has become the ideal for every modern family. In Pakistan, most of the people use Android phones for improving the living style but overall environment is not secure and safe. Therefore, it is a great need of automation systems. The GSM based android software application along with the security system is developed and presented in this paper. The developed application is useful enough for users including handicapped persons.

II. Literature Review

Many research contributions have been published pertaining to the home automation and automatic security systems. The fundamental information regarding the home automation and security systems implemented with Arduino and GSM technology are described and presented by Kaur [5]. Some researchers used Bluetooth technology in networking environment as well as automation systems for instance; Sriskanthan [6] developed an application for home automation using Bluetooth technology.

In previous past, home automation systems are ambiguous and complicated due to the system hardware but nowadays these systems are used by many people across the world with modern technologies. Touch screen based home automation system is developed by Wagh [7] using GSM and Zig-Bee. The GSM technology is also used by Singh [8] in the developed appliance system. The android application having low cost and provide switching services is presented and with object oriented programming language by Pawar [9]. The Internet of Things was used in the proposed system for controlling and monitoring the various appliances. GSM is widely used in such kind of application that controls the appliances. In Pakistan most of the offices, business platforms and educational institutes have not any appliances security system that provide facility to control the devices through-out the smart phone applications. Thus, no work has been found in Pakistan from the side of security control system.

The security of homes is mandatory in countries like Pakistan. The survey was conducted by Chitnis [10] from the peoples having different background for the awareness of the automatic home automation system and its significance particularly in terms of security. The home security system is developed and presented by Mali [11] using motion sensor and PIR sensor. It is noted that Arduino board and GSM is used for data processing and messaging. Moreover, low power Bluetooth protocol is used in automation system by Chandra [12] with suitable authentication for correct person. On the other hand, PIC based remote control system is presented by Erol [13] for intelligent homes. This system is electrically and optically isolated system that is much secure. The Pin-Check algorithm is used in developed system in order to enhance and improve the security.

III. System Design and Execution Process

The core purpose of this project is to develop software application for home automation and security control system using different hardware components. Various GSM Modules having different frequencies are available as shown in "Fig. 1" but GSM SIM900A Module is selected in this project. This Module is used because of the availability, coverage and security and it is widely used for establishing the connections where Internet access is not possible. The server uses AT commands to communicate with the GSM modem [4] for controlling the SMS and send it to the Arduino for further process.

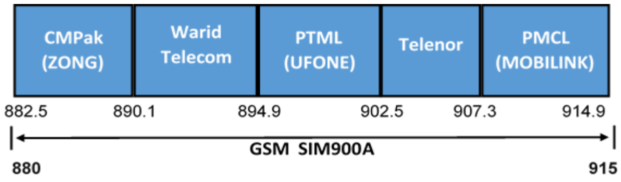


Fig. 1. GSM Module Frequencies

During the literature review, it has been found that Arduino board is used as the controller to interface the appliances [3]. When the embedded systems start, user first receives a message that the system is ready for process. It uses certain peripheral drivers and relays for successful communication and controlling the load of appliances. The touch screen smart phone is used to handle the application using the GUI. The developed application automatically generates SMS messages based on the user commands and sends it to the GSM modem attached to the Arduino. This allows the user to control the selected home appliances as well as security system. The complete system architecture of the developed software project is depicted in "Fig. 2".

Furthermore, for understanding and easy accessing there is a GUI for controlling the different devices such as bulb saver, fan etc. The list of various functions is available in the core screen of the application. The user can select any function among the available list for controlling the device. After that user can see the action performed on the selected device. Now if user wants to enable or disable the particular device of the appliances then application give a facility to do as per the requirement and wish of the user. Now the function of GSM is begins, the GSM module decodes the received messages via SMS and performs the given commands. It is noted here that, the SMS depends on the used networks and there is a possibility of late delivery of the message due to the availability of the signals. The flow chart of the complete execution process of developed software application is shown in "Fig 3". The flow chart is taken from [14] except the LAN because GSM is used in this project.

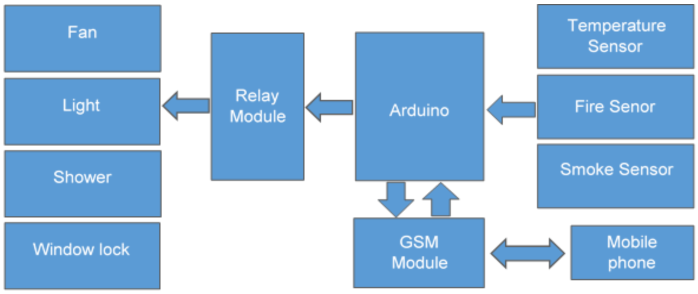


Fig. 2. System Architecture of Developed Software Application

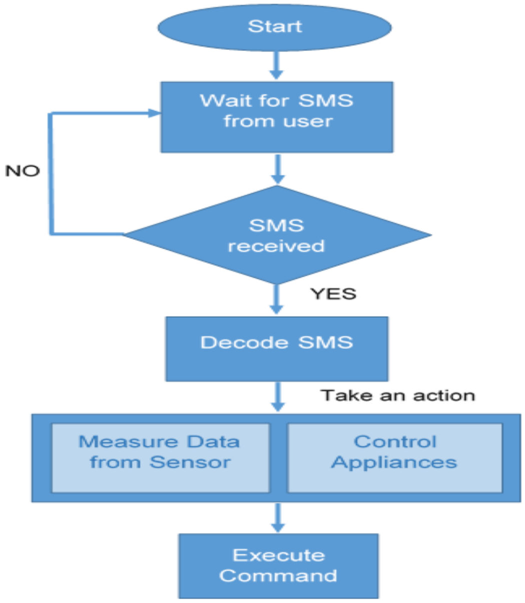


Fig. 3. Execution Process of Developed Software Application

The developed application consists of four components: Light controlling, Door controlling, Fire & Smoke detection and Temperature sensing. The application has ability to automatically activate the alarm when system detects any symptom of smoke, gas or fire. If fire detected then the water shower is activated. Noted that, if user is not present in his home then he can watch the live streaming through Internet. One of the ability of the developed application, user can control multiple appliances concurrently because the system is able to verify the status of the appliances simultaneously. For this process, security camera is connected to the internet and sending live video streaming to a domain like YouTube. Mobile phone facilitates to watch that streaming which is available on to the domain with corresponding URL. The accessing process of security camera is depicted in "Fig.4".

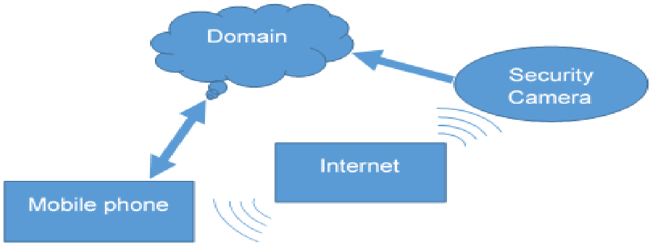


Fig. 4. Accessing Process of Security Camera

IV. Results and Discussion

The developed software application with complete automation system is individually tested in three different cities of province Sindh. The selected cities are Karachi, Sukkur and Khairpur, these cities are selected due to the performance evaluation of developed application and concerned hardware. For the analysis and further comparison, five telecommunication companies i.e. ZONG, Warid, Ufone, Telenor and Moblink are selected which are less or more works on these selected cities. Moreover, three different places are selected in each city for testing the frequency of the GSM Module. The aim of this analysis and comparison is to explore the availability of the signals with certain ranges and the performance of the developed software application.

A. *Karachi*

The developed system is practically tested in three different areas of Karachi i.e. Gulshan-e-Iqba, Cant Station and Ghulistan –e- Johar. It is noted that developed system is tested in ten different locations of each area of Karachi city and calculated the performance in percentage in terms of the availability of the signals and functionality of the application. The average performance accuracy with ZONG is calculated 55%, in selected three areas 81.6% is calculated with Ufone, 65% and 70% is achieved with Telenor and Moblink respectively. The detailed calculated results of developed application at Karachi are shown in Table I and calculated results are graphical represented in “Fig.5” for analysis and comparison.

The calculated results proved that Ufone is better than the other telecommunication networks because 81.6% performance is achieved with this network. On the other, inferior results are received with ZONG. The results of Telenor and Moblink are also at acceptable level.

TABLE I. Calculated Results of Developed Application at Karachi

City: Karachi				
Selected Network	**Selected Areas**			**Mean (%)**
	Gulshan - e-Iqbal (%)	**Cant Station (%)**	**Gulistan e Johar (%)**	
ZONG (882.5—890.1 MHz)	40	65	60	55
Ufone (894.9 - 902.5 MHz)	70	85	90	81.6
Telenor (902.5 – 907.3 MHz)	50	80	65	65
Moblink (907.3 - 914.9 MHz)	55	70	85	70

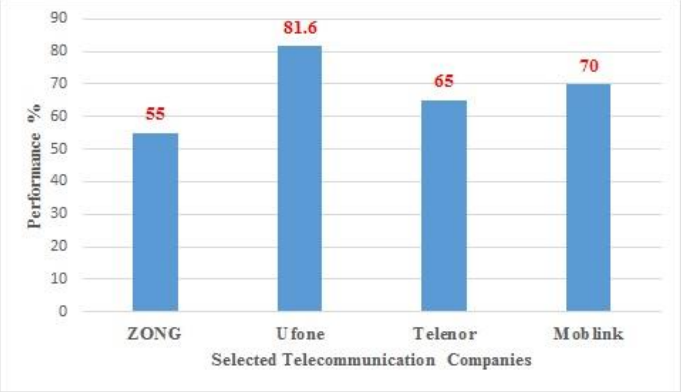


Fig. 5. Calculated Performance with Telecommunication Companies at Karachi

B. *Sukkur*

For testing the developed software project at Sukkur, three areas such as Airport Road, Race Course Road and Old Sukkur. Ten different locations of Sukkur were selected for evaluating the performance of selected telecommunication networks using our developed project. We have achieved the better accuracy after testing the proposed system. Thus, the average performance accuracy with ZONG is calculated 90.6%. The performance accuracy of 81.6% is calculated with Ufone, 76% is achieved with Telenor and 85% is achieved with Moblink. The statistics of calculated results at Sukkur city is given in Table II and calculated results are graphical represented in “Fig.6” for analysis and comparison.

TABLE II. CALCULATED RESULTS OF DEVELOPED APPLICATION AT SUKKUR

City: Sukkur				
Selected Network	Selected Areas			Mean (%)
	Airport Road (%)	Race Course Road (%)	Old Sukkur (%)	
ZONG (882.5—890.1 MHz)	90	95	87	90.6
Ufone (894.9 - 902.5 MHz)	70	85	90	81.6
Telenor (902.5 – 907.3 MHz)	75	80	73	76
Moblink (907.3 - 914.9 MHz)	85	90	80	85

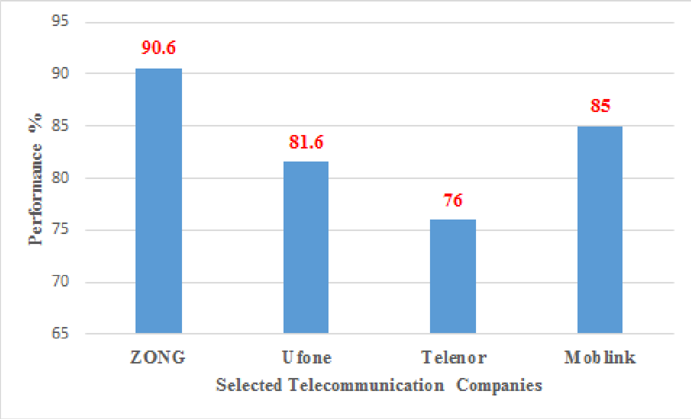


Fig. 6. Calculated Performance with Telecommunication Companies at Sukkur

As fig.6 shows that ZONG is more suitable communication network at Sukkur because high performance is calculated with this network which is 90.6%. With small variations, other three networks are also performed well. However, poor results are achieved with Telenor which is 76%.

C. Khairpur

The developed software application is also tested in various locations of Khairpur. For this, three areas of Khairpur i.e. SALU Khairpur, Station Road and Jillani Muhalla are selected. The performance of the system is calculated using the formula of mean. The performance accuracy with ZONG is calculated 45%, in selected three areas 75.3% is calculated with Ufone, 68% is calculated with Telenor and 71.6% is calculated with Moblink. The statistical information of calculated results is shown in Table III and calculated results are also graphical represented in “Fig.7” for comparison.

TABLE III. CALCULATED RESULTS OF DEVELOPED APPLICATION AT KHAIRPUR

City: Khairpur				
Selected Network	Selected Areas			Mean (%)
	SALU Khairpur (%)	Station Road(%)	Jillani Muhala(%)	
ZONG (882.5—890.1 MHz)	30	45	60	45
Ufone (894.9 - 902.5 MHz)	55	84	87	75.3
Telenor (902.5 – 907.3 MHz)	64	67	73	68
Moblink (907.3 - 914.9 MHz)	75	72	68	71.6

On the basis of received results, it is concluded that Ufone telecommunication network is more suitable than others because 75.3% performance is achieved with this network. The lesser results are received with ZONG. The results of Telenor and Moblink are also at acceptable level.

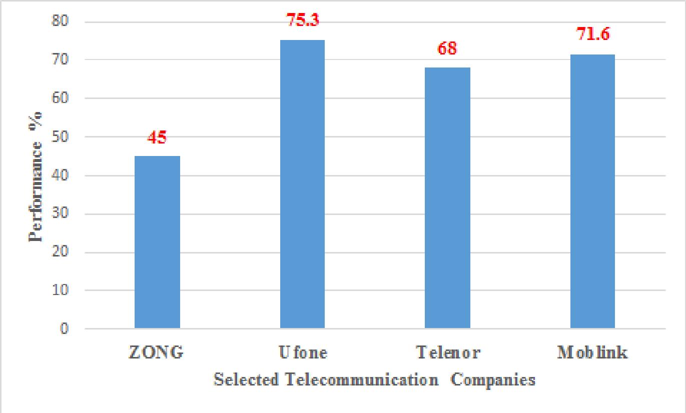


Fig. 7. Calculated Performance with Telecommunication Companies at Khairpur

V. CONCLUSION

The automation applications permit people to control the appliances used in their homes, offices, hospitals etc. It provides security system with cameras for controlling and monitoring activities around the home. This paper presented the system architecture and the calculated results which are testing at different areas and locations of Karachi, Sukkur and Khairpur with different telecommunication networks used in Pakistan. Authors calculated performance of 81.6% in Karachi using Ufone, 90.6% is calculated with ZONG at Sukkur and 75.3% is received with Ufone at Khairpur. On the basis of calculated results, it is proved that Ufone is better than the other selected telecommunication networks. The outcome of this study will be helpful for the students as well as researchers who want to put their efforts towards the designing and development of automation & security systems. The future scope of proposed system will led to implement rigid security for schools and other educational environment.

References

[1] C. A. Jose, R. Malekian, “Smart home automation security: A literature review”, Smart Computing Review, Vol. 5, No. 4, Pp. 269-285, 2015.

[2] S. B. Priya, R. Geethamani, “Design and implementation of home automation using power electronic switches”, International Journal of Advanced Information and Communication Technology, Vol. 2, Issue 12, Pp. 1127-1129, 2016.

[3] D. Javale, M. Mohsin, S. Nandanwar, M. Shingate, “Home automation and security system using android adk”, International Journal of Electronics Communication and Computer Technology, Vol. 3, Issue 2, Pp. 382-385, 2013.

[4] S. Palaniappan, N. Hariharan, T. N. Kesh, S. Vidhyalakshmi, Angel, S. Deborah, “Home automation system- A study”, International Journal of Computer Applications, Vol.116, No. 11, Pp. 11-18, 2015.

[5] S. Kaur, R. Singh, N. Khairwal, P. Jain, “Home automation and security system”, Advanced Computational Intelligence: An International Journal, Vol. 3, No. 3, Pp. 17-23, 2016.

[6] N. Sriskanthan, A. Tan, A. Karande, “Bluetooth based home automation system”, Microprocessors and Microsystems, Vol. 26, Pp. 281–289, 2002.

[7] M. Wagh, V. Gadhari, H. Sonawane, S. Shelar, R. Mahale, “Touch screen based home automation system”, International Research Journal of Engineering and Technology, Vol. 3, Issue 3, Pp.1530-1531, 2016.

[8] P. Singh, K. Chotalia, S. Pingale, S. Kadam, “A review paper on smart gsm based home automation system”, International Research Journal of Engineering and Technology, Vol. 3 Issue 4, Pp.1838-1843, 2016.

[9] N. P. Pawar, S. Ramachandran, P. N. Singh, V. V. Wagh, “A survey on internet of things based home automation system”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, Pp. 76-81, 2016.

[10] S. Chitnis, N. Deshpande, A. Shaligram, “An investigative study for smart home security: Issues, challenges and countermeasures”, Wireless Sensor Network, Vol. 8, Pp. 61-68, 2016.

[11] V. Mali, A. Gorasia, M. Patil, P. S. Wawage, “Home automation and security using arduino microcontroller”, International Journal of Research in Advent Technology, Special Issue National Conference, Pp. 214-217, 2016.

[12] G. P. Chandra, S. Venkateswarao, “Ad-Hoc low powered 802.15.1 protocol based automation system for residence using mobile devices”, International Journal of Computer Science & Technology, Vol. 2, No. 1, Pp. 93-96, 2011.

[13] Y. Erol, H. Balik, S. Inal, D. Karabulut, “Safe and secure pic based remote control application for intelligent home”, International Journal of Computer Science and Network Security, Vol.7, No.5, Pp. 179-182, 2007.

[14] P. P. Kumar, T. G. Vasu, “Home automation & security system using arduino android adk”, International Journal of Emerging Trends in Engineering Research, Vol. 3, No. 6, Pp. 190-194, 2015.

ROI-based Compression on Radiological Image by Urdhva-Tiryagbhyam and DWT Over FPGA

Suma
Research Scholar
Vidya Vikas Institute of Technology
Mysore, India

V. Sridhar
Principal, PES College of Engineering
Mandya
Karnataka, India

***Abstract*—The area of radiological image compression has not yet met its potential solution. After reviewing the existing mechanism of compression, it was found that majority of the existing techniques suffers from significant pitfalls e.g. more usage of transformation schemes, more resource utilization, delay, less focus on FPGA performance enhancement, extremely less emphasis on Vedic-multipliers. Hence, this paper presents an analytical modelling of ROI (Region of Interest)-based radiological image compression that applies Vedic Multiplier Urdhva-Tiryagbhyam to enhance the performance of coding using Discrete Wavelet Transform (DWT). The study outcome was implemented in Matlab and multiple test bed of FPGA devices (Virtex 4 FX100 -12 FF1152 and Spartan 3 XC400-5TQ144) and assessed using both visual and numerical outcomes to find that proposed system excel better performance in comparison to recently existing techniques.**

Keywords—Radiological Image Compression; Discrete Wavelet Transform; FPGA; Lifting Scheme

I. Introduction

With the advancement of radiological image processing, the area of healthcare and the diagnostic sector has been imparted with various means to visualize the diseases with higher accuracies. Various kinds of radiological imaging systems e.g. Ultrasound, Positron Emission Tomography, Computed Tomography scans, Magnetic Resonance Imaging, have assisted radiologist and physician to investigate the disease very closely [1]. With the evolution of cloud, storage is never a problem and hence the adoption of cloud or any pervasive computing significantly assists in storing such radiological images. However, there is a darker side of this story even at present about radiological image. Normally, radiological images are very different from other types of images in the form of information contents. A normal MRI image can be around 5-6 MB, whereas the size can further increase [2]. Although storage is not a problem, a problem arises in particular applications e.g. telemedicine, robotic surgery, etc. [3][4][5]. In such applications, it is required that a bigger size of the image may need to be transferred from one to another end of the world with least delay as possible [6]. There are many network parameters e.g. traffic congestion, channel capacity, interference, noise that always affects the transmission [7], which is less likely to be controlled. However, a better transmission of a radiological image will also require that image retains its maximum signal quality which will degrade in the process of transmission. The solution of all this is an effective compression algorithm. An effective compression algorithm not only reduces the size but also ensures to retain maximum signal quality while reconstructing it at the end receiver [8]. At present, there are various lossless and lossy compression schemes [9] to accomplish maximized compression ratio, but the biggest challenge of lossy schemes is to recover the original data. On the other hand, usage of lossless compression schemes minimizes the compression ratio although it has the potential to recover the complete data. The significant problem is that lossless scheme is the only alternate solution in radiological image compression as it can’t allow losing any forms of clinical information at any cost. Another problem is that bigger images e.g. MRI images has a bigger background which is not required to be processed at all as it occupies half of proportion size of the entire image resulting in the expensive matter. Moreover, such null background also has artifacts as well as noise during the process of acquisition of an image. Hence, implementation of denoising techniques further degrades the signal quality [10]. At present, there is an increasing trend of using JPEG2000 standards owing to its supportability of advanced characteristics of radiological image processing. It is found to have an optimal efficiency of coding in comparison to conventional compression scheme [11]. JPEG2000 also has higher supportable features of Discrete Wavelet Transform (DWT) along with various arithmetical schemes of coding [12]. There is an increasing attention from the research community towards using DWT-based scheme and its associated architecture. It was also found that FPGA has enough potential to hone the potential features of DWT. There have been a various research attempts on using FPGA and DWT together in image processing [13]. All these implementations suffer from certain flaws e.g. more dependencies on control signals, inferior hardware utilization, more latency, increasing demands of registers in hardware design, etc. This paper introduces a technique that jointly uses Vedic Multiplier, FPGA (Field Programmable Gate Array), lifting scheme, and DWT (Discrete Wavelet Transform) to overcome the problems. Section II discusses the existing research work towards the topic followed by a briefing of problem identification in Section III. Section IV introduces the adopted methodology of proposed system followed by a discussion of algorithms in Section V. Result accomplished from the study is discussed in section VI followed by a summary of the work in conclusion under Section VII.

II. Related Work

This section discusses the techniques adopted for radiological image compression during the year 2010-2015.

Discrete Cosine Transform is one of the most frequently used compression algorithm for images. The work carried out by Dhandapani and Ramachandran [14] have presented an 8x8 transformation matrix needed by an added and thereby skipping shift and multiplication using FPGA. Although, the outcome was found to exhibit power effectiveness with lesser delay compared to other existing system but it doesn't prove computational efficiency. Another frequently used technique to perform compression is DWT and SPIHT (Set-Partitioning in Hierarchical Trees). Fang et al. [15] have introduced a unique interpolation-based scheme using Lagrangian theory in to show better supportability of prediction depending upon local features. The study outcome was found to minimize execution time along with the length of coding bits but was not found to offer better signal quality of reconstructed image. Usage of DWT for radiological image compression was also seen in the work of Govindan and Sanile [16], where it was used for truncating the sub-bands in ultrasonic images. The authors have also applied interpolation technique using Fourier transform to obtain enhanced signal quality. The study has used both spatial and temporal correlational properties in to perform compression. The study outcome was assessed to find better signal quality (measured by PSNR), correlational coefficient, and quality of reconstruction, computational speed, and resource utilization. However, the technique was not found benchmarked. Same authors [17] have extended the similar work using system-on-chip platform using OpenCL language over GPU. Medical images are not only limited to radiological images, but there are also other forms of it. One of such form is DNA microarray, which is frequently used in genetic engineering. Cabronero et al. [18] have particularly addressed the problem of compressing such medical image of DNA microarray using a quantization-based approach. The contribution of the work is to restrict the relative error due to quantization. On increasing bitrates, the study was found to have better error control. Rehman et al. [19] have presented a discussion on bi-orthogonal transform. The technique takes the input image and converts it to macro blocks which are then further transformed into blocks and pixels. The authors have used JPEG XR to perform compression with lowered memory usage. Yoon et al. [20] have applied L-fold down sampling to minimize the coefficients of filters and data rates.

Apart from transform-based implementation techniques, there is a frequent usage of FPGA-based schemes too for performing compression. Ahmad et al. [21] have introduced a three-dimensional daubechies for compressing the medical image. The three dimensionalities are achieved by considering two transpose buffers and one-dimensional Daubechies. The study outcome was found with lower consumption of power but without benchmarking. Anjaneyulu and Krishna [22] have used conventional DWT as well as SPIHT on FPGA. However, this work is a replica of original work carried out by Fang et al. [15] and doesn't show many novel features in work. Usage of FPGA was also advocated by Botella et al. [23]. In existing system, FPGA was used in a different way too. For example, the work carried out by Li et al. [24] has developed a synthesis tool using FPGA. However, the study didn't provide enough evidence to claim its effectiveness. Nagabhushanam et al. [25] has also used FPGA-based approach and DWT-based scheme to perform image compression. The authors have addressed the complexity of DWT by incorporating the enhanced version of DWT using distributive arithmetic. The study outcome was found with lowered latency and increased throughput regarding clock cycles. However, the outcome didn't studied computational complexity. Wu et al. [26] have applied SPIHT along with FPGA to perform image compression. The technique incorporates parallelism over SPIHT algorithm to enhance the processing capacity. The basic technique is to retain the maximized PSNR and minimize the storage demands. However, the technique is not cost effective if the dataset is changed to complex medical image. Xuesen et al. [27] have designed an experimental test-bed that uses FPGA for investigating the acquired data from the medical image.

Among all these techniques, the performance of the system can be greatly enhanced if a suitable multiplier is applied to maintain a balance between compression and quality of the reconstructed image. In this regards, Vedic-based multipliers designs have also been researched to some extent. Most recently, Arish and Sharma [28] have presented a design of multiplier for floating point that significantly controls both power dissipation and delay. The authors have used two Vedic-based schemes e.g. Urdhva-Tiryagbhyam algorithm and Karatsuba algorithm to deploy binary multiplier of the unsigned type for the purpose of carrying out mantissa multiplication. An exactly similar version of work is also carried out by Kodali et al. [29] in the same year. Usage of Vedic mathematics was found in work carried out by Pohokar et al. [30]. The work was carried out in FPGA and outcomes are found with reduced delay and memory demands compared to the traditional multiplier. Nearly similar work was also carried out by Sharma and Goyal [31] most recently only with a difference that actual work of Pohokar [30] was implemented in FPGA in 2015 and same work was also published by Sharma and Goyal [31] in H-Spice in 2016. Singh and Sasamal [32] have presented a study where binary Vedic multiplier is implemented over cadence tool using adiabatic logic. Usage of another Vedic sutra called as Calana Kalanabhyam was seen in the work of Verma et al. [33] where the authors have used it alongside with FPGA to make energy-efficient sutra. Vijayan et al. [34] have presented a Vedic multiplier of 8-bits in FPGA. Hence, it can be seen that there are quite a lot of study being carried out on the topic of image compression using DCT, Fourier transforms, FPGA, Vedic multiplier, SPIHT, etc. All the studies have significant points to learn as well as pitfalls too. The Vedic multiplier, although being an older concept, has not been much explored in the area of medical image compression. The next section elaborates about the problems being identified from the existing studies.

III. Problem Identification

This section discusses the problem identification after reviewing the literature from the existing system:

- **More inclination towards transformation-based schemes**: It has also been seen that transformation based schemes (e.g. DWT, DCT, SPIHT) are more used for performing compression. However, there is some potential trade-off in using such schemes which were completely not highlighted in any of the research work till date. Although wavelet-based schemes support compression of both lossy and lossless image, unfortunately there are many modalities that are found not support the generation of compressed objects in JPEG2000 [35]. Existing techniques doesn't addressed a problem that compressed JPEG2000 objects are not supported by Picture Archiving And Communication System (PACS) that forces converting the image to some other formats while transmission. Moreover, usage of the wavelet-based scheme includes higher computational cost [35].

- **Less focus on enhancing FPGA**: At present, there are many FPGA-based schemes using radiological image processing. However, the biggest research gap is FPGA is just used as a platform for synthesis. For example, the work carried out by various researchers [24][25][26] have used FPGA using an image. However, it was totally ignored that image conversion module over various devices of FPGA includes various steps that are a computationally intensive process. There are also studies that have used FPGA and DWT for image compression using distributed arithmetic which includes increasing number of shift registers on LUT. The techniques using DWT and FPGA cannot optimize the entropy encoding.
- **Less Emphasis on Vedic Multipliers**: There are only 2 transaction papers and 155 conference papers on Vedic-based approaches published during 2010-2016 in IEEE Xplore. However, about radiological image compression, there exist only two conference papers that have implemented Vedic multiplier during last 5 years [36][37]. This itself is one of the potential evidence that there has been quite a less emphasis on standard research work towards realizing the potential characteristics of Vedic-based approaches in medical image compression.

Apart from the above-mentioned problems, it was also found that existing techniques suffers from certain common problem viz. i) lack of benchmarking, ii) inappropriate architecture usage leading to poor process of image reconstruction, ii) considers the entire image for compression leading to more bandwidth utilization as well as computational resource utilization, etc. Moreover, the potential of the Vedic multiplier, DWT and FPGA are less explored jointly. Hence, the problem statement can be stated as "*It is a challenging task to develop a faster radiological image compression considering the joint potential of DWT, FPGA, and Vedic Multipliers*." The next section about the methodology adopted in to address the identified problems.

IV. Proposed Methodology

The design and implementation of the proposed work are carried out using analytical-based methodology. The present work is a continuation of our prior works [36][38][39]. The proposed system implements Vedic multiplier called as *Urdhva-Tiryagbhyam*. The beneficial point of this multiplier is that it can concurrently handle addition and fractional product generation. This feature gives more edge to parallel processing that can significantly reduce the delay time. It also makes it highly suitable for performing binary multiplication. Fig.1 highlights the architecture designed for the proposed system.

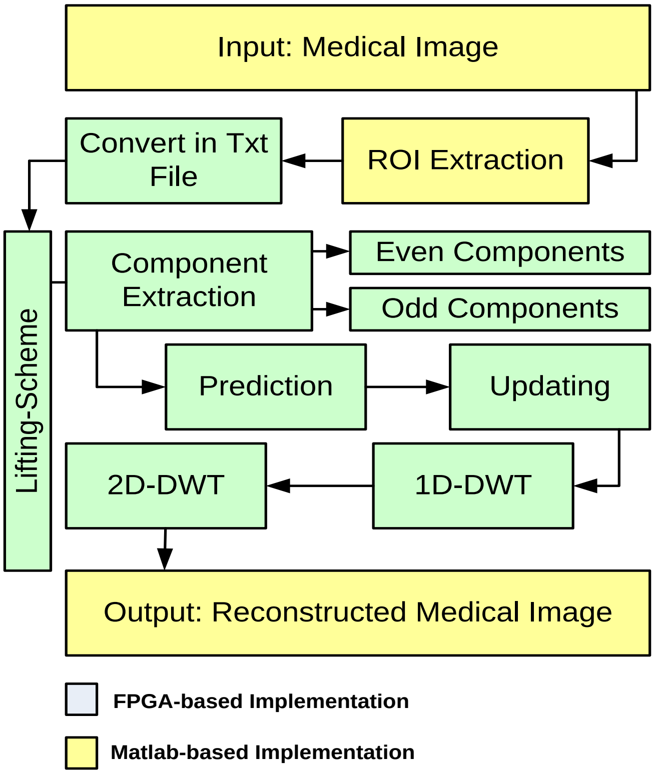


Fig. 1. Proposed Methodology

The proposed system uses Region-of-Interest for taking the input of radiological image, which means the proposed algorithm is only applicable to ROI and not the complete radiological image. It than converts the image into a binary text file, so that it can be processed effectively in FPGA. The next step is to apply lifting scheme which has its inherent advantages over eliminating redundancies. It works by classifying the ROI data to even and odd samples further followed by prediction and updating operation. Prediction process assists in obtaining approximated coefficients while updating operation assists in obtaining detailed coefficient values. This step is carried out to resist lossy data and to obtained lossless data during the entire process of radiological image compression. Finally, it generates high and loss pass components of one-dimensional DWT which is further processed back to obtain two-dimensional DWT in the form of a reconstructed image. The elaborated discussion of the algorithm of proposed system is carried out in next section.

V. Algorithm Implementation

The main purpose of this algorithm is to perform radiological image compression using the Vedic multiplier. The proposed system uses radiological image database of [40] in algorithm implementation. In spite of compressing the entire image, we choose to compress only the Region-of-Interest (ROI). The advantage is faster processing, less allocation of computational resources, and retention of the image portion of clinical importance. The proposed system uses Matlab to take the input of radiological image and then convert it into a text file (a form of a hexadecimal number), which is further subjected to FPGA (Line-2). The entire algorithm implementation is carried out over FPGA itself. It is already known that a filtering operation can be carried out over DWT that can be represented as,

$$M(i) = \prod_{j=1}^{m} \begin{pmatrix} TL_e(i) & TH_e(i) \\ TL_o(i) & TH_o(i) \end{pmatrix} \tag{1}$$

In the above equation, *TL*(*i*) and *TL*(*i*) corresponds to transfer function of low as well as high pass filter. It also means that $TL_e(i)$ and $TH_e(i)$ corresponds to even components while $TL_o(i)$ and $TH_o(i)$ represents odd components. The better representation of the transform function could be,

$$(\varphi(i) \quad \theta(i)) = (x_e(i) \quad \frac{1}{z} x_o(i)) M(i) \tag{2}$$

In the above equation, the variables $\varphi(i)$ and $\theta(i)$ represents both low and high pass components that have been filtered from its input signal $x(i)$. The proposed system also uses lifting scheme that can further factorize the matrix representation giving better feasibility to enhancing the capability of the processor. Another advantage of applying lifting scheme will be to eliminate the redundancies. As per Fig.2, even clocking will be functional on the registers position on top while odd clocking will be operational for the registers positioned on the bottom. Using a clock cycle of a unit pixel, the ROI image data is fed in serial order to classify the data as even (c_e) and odd components (c_o) (Line-3). This operation will lead to the implementation of lifting scheme that further results in low pass components (LP_c) and high pass components (HP_c) (Line-4). A unit of low pass coefficient, as well as high pass coefficient will be generated for a single unit of ROI image.

Algorithm for Compressing Radiological Image

Input: I_{roi} (image), c_e / c_o (even and odd component), LP_c / HP_c (Low and High pass components)

Output: I_{recon} (reconstructed image)

```
Start
1. init ()
2. read $I_{roi}$, $I_{roi}$→txt()
3. [α, β]=classify(txt)→[$c_e$, $c_o$]
4. get()=[$LP_c$, $HP_c$]
5. get()→($x_5$[n], $x_6$[n])→1D DWT
6. Apply Urdhva-Tiryagbhyam
7. $I_{recon}$=get(txt)← 2D DWT
End
```

The generated low pass component and high pass components are then identified. The proposed system also makes use of Daubechies 9/7 filters. The advantage will be an enhancement in the compression performance will be retained to maximum level with proper control over computational complexity. This implementation policy can be seen in Fig.1, where the ROI image data is forwarded through numerous steps. The numbers of transformed coefficients are retained to be similar as that of original one owing to sub-sampling. The system than processes the data to obtain detailed (x_5(i)) and approximation coefficients (x_6(i)) of one-dimensional DWT (Line-5). Implemented over FPGA using Verilog, the coefficient outcome was obtained.

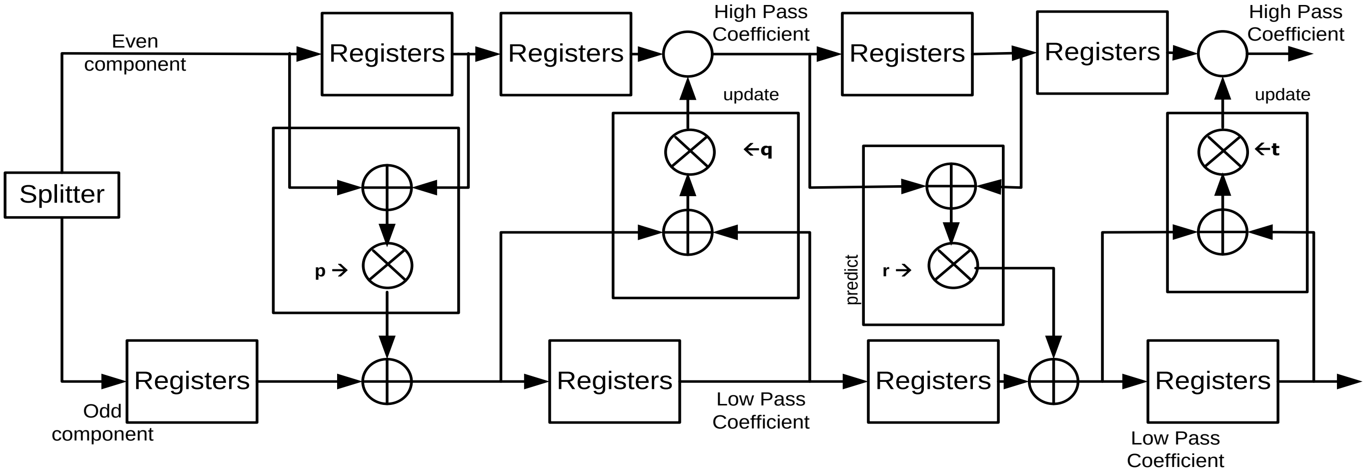


Fig. 2. Algorithm Operations in DWT

The next part of the algorithm implementation is to apply Vedic multiplier design using hardware-based architectural approach (Line-6). The implemented design scheme of the Vedic multiplier was shown in Fig.3.

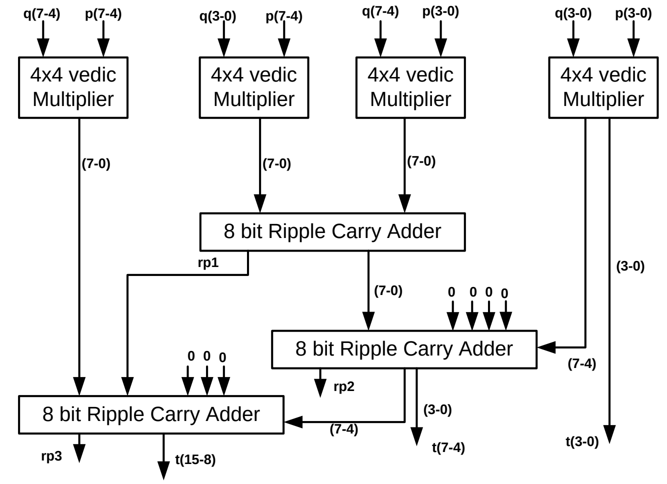


Fig. 3. 8x8 Vedic Multiplier Design

The proposed technique has been testified for both 2x2 as well as a 4x4 bit of Vedic multiplier to design the hardware architecture. The system uses Urdhva Tiryagbhyam sutra which is meant for vertical and crosswise multiplication of dual-number of binary origin. Initially a 4x4 bit of Vedic multiplier has been designed, which further uses 8 bit of ripple carry added to formulate an 8x8 bit of Vedic multiplier. The entire process starts with 2 bit numbers say P and Q to for 2x2 bit of Vedic multiplier using vertical and crosswise multiplication of least significant bits. It is then enhanced with 4x4 bit of Vedic multiplier considering two-bit at a time in 2-bit block of the multiplier. Although, it can reduce delay, the delay performance can be further enhanced by considering four 4x4 bit of Vedic multiplier and three 8 bit of ripple carry adder. Consider P=P_7 P_6 P_5 P_4 P_3 P_2 P_1 P_0 and Q=Q_7 Q_6 Q_5 Q_4 Q_3 Q_2 Q_1 Q_0. Therefore, according to the concept of Vedic-based multiplication, it will result in 16 bits as T_{15} T_{14} T_{13} T_{12} T_{11} T_{10} T_9 T_8 T_7 T_6 T_5 T_4 T_3 T_2 T_1 T_0. Dividing the bits P and Q will result in further decomposition into a pair of minimum 4 bits of sub-bits e.g. P_{high} and P_{low}. Similar generation will be yield by Q also as Q_{high} and $Q_{low.}$ The further processing of multiplication is carried out by utilizing 4 bits of multiplier blocks and considering 4 bits at a single instance, which is by the Vedic multiplier theorem of Urdhva Tiryagbhyam sutra. Finally, the resultant is accomplished from the addition of the multipliers of 4x4 bit output. In this entire process, three ripple carrier adders of 8 bits are also used as shown in Fig.3. The final-outcome from the FPGA is than fed to Matlab to obtain a reconstructed image (Line-7). Hence, the proposed system wisely utilizes the Vedic multiplier to enhance the compression performance over radiological images. The next section discusses the results being accomplished by the implementation of the algorithm discussed in this section.

VI. Result Discussion

The proposed system was implemented on 32 bit windows system with normal 4 GB of RAM and core-i5 processor. The implementation is carried out over two types of FPGA devices i.e. i) Virtex 4 FX100 -12 FF1152 and ii) Spartan 3 XC400-5TQ144. The formation of the one-dimensional DWT with Vedic multiplier using Xilinx and Verilog is shown in Fig.4

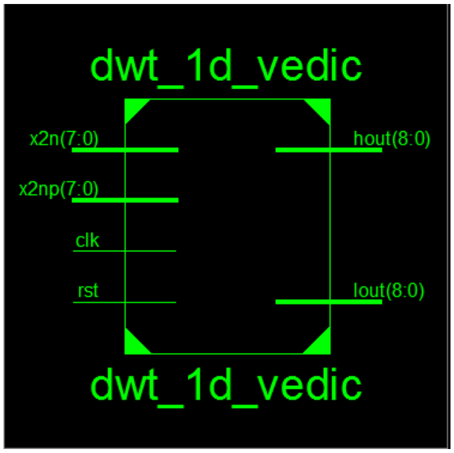


Fig. 4. Module of 1D_DWT

Upon receiving the input data, it is classified in even and odd components which are then reposited in registers. The registers will have a null value when there is higher reset, but during lower reset, the registers classify the input ROI image in the form of odd and even components. The lifting schemes enable to read the input ROI data to 16 clock cycles whereas the resultant outcome will have both high and low pass components according to the algorithm steps (one-dimensional-DWT). The module shown in Fig.4 takes the input of pixel values, which categorizes the values in odd and even components and develops a matrix where it is stored. This operation is then followed by the Vedic multiplier to further leverage the compression process.

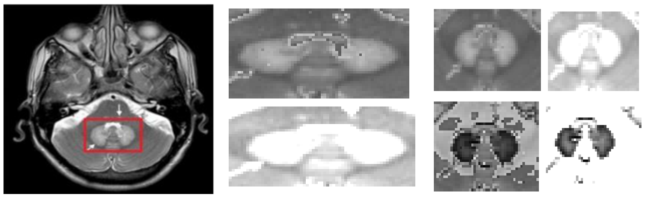


Fig. 5. (a) Input image (b) 1D-DWT results (c) 2D-DWT results

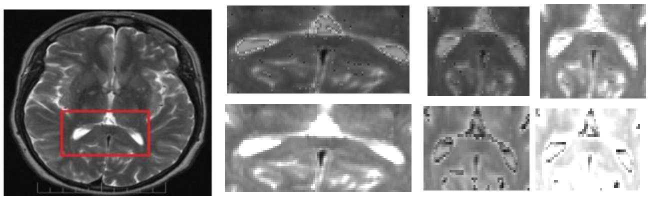
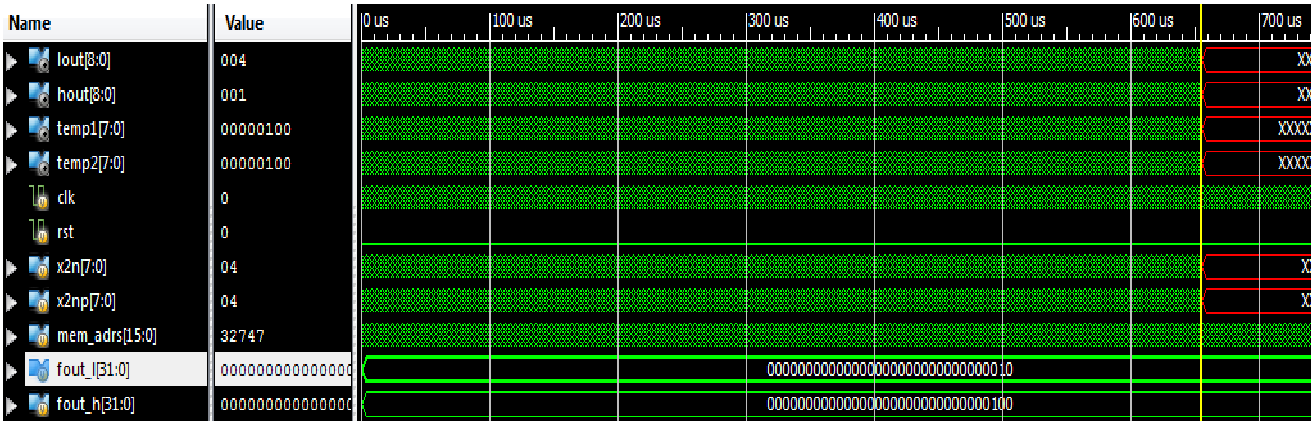


Fig. 6. (a) Input image (b) 1D-DWT results (c) 2D-DWT results

Fig. 5-6 shows the visual outcomes of the two sample radiological image whose ROI is considered for performing compression operation using proposed system. Both the visual outcomes show accomplishment of one-dimensional DWT outcomes, which is then followed by two-dimensional DWT outcomes. A closer look at the reconstructed two-dimensional DWT image shows better perceptibility of the processed radiological image after applying compression.



Name	Value
lout[8:0]	004
hout[8:0]	001
temp1[7:0]	00000100
temp2[7:0]	00000100
clk	0
rst	0
x2n[7:0]	04
x2np[7:0]	04
mem_adrs[15:0]	32747
fout_l[31:0]	000000000000000
fout_h[31:0]	000000000000000

Fig. 7. Simulation results of 2D-DWT

Fig.7 highlights the simulation outcome of two dimensional DWT in FPGA. For effective analysis, the design of the proposed system is compared with the similar type of work presented by Todkar [41] and Mhamunkar [42] most recently. The standard performance parameters are retained e.g. logic utilization, number of slices, number of slice flip flops, number of 4 unit LUT, and number of bonded IOBs.

The work carried out by Todkar [41] was mainly focused on enhancing the operation of two dimensional DWT over VLSI-based architecture. The authors have used 9/7 filter design as the lifting scheme that is quite similar to us. The technique has minimized the higher dependencies of registers to have better control over a delay. The authors have amended the design of DWT, transpose unit, processing elements, and mechanism of prediction and update. The implementation was carried out using 16 multipliers, 5 input buffers, 54 registers, 24 adders, and 11 transposition registers with lifting based DWT scheme. The numerical outcome was presented in Table 1 using FPGA device (Virtex 4 FX100 -12 FF1152).

TABLE I. NUMERICAL COMPARATIVE ANALYSIS-I

		Proposed Design		Todkar's approach [41]	
Logic Utilization	Available	Used	Utilization	Used	Utilization
Number of Slices	42176	770	1%	878	2%
Number of Slice Flip Flops	84352	295	0%	1072	1%
Number of 4 input LUTs	84352	1394	1%	1263	1%
Number of bonded IOBs	576	36	6%	NA	6%
Number of GCLKs	32	1	3%	1	3%

The first significant improvement can be seen is the lowered dependencies on the utilization of slices by proposed system and it almost doesn't use any flip flops. The complete processing time of the algorithm for proposed system is found to be 47% improved as compared to Todkar's approach [41] with less computational complexity

TABLE II. NUMERICAL COMPARATIVE ANALYSIS-I

		Proposed Design		Mhamunkar's approach [42]	
Logic Utilization	Available	Used	Utilization	Used	Utilization
Number of Slices	1920	742	38%	1880	97%
Number of Slice Flip Flops	3840	295	7%	2118	55%
Number of 4 input LUTs	3840	1340	34%	2971	77%
Number of bonded IOBs	97	36	37%	62	63%
Number of GCLKs	8	1	12%	4	50%

The above Table 2 shows the numerical analysis of proposed system with that of work carried out by Mhamunkar [42]. The author have used frequently used image compression algorithm i.e. SPIHT with an objective to maintain balance between compression and image quality over FPGA. The algorithm processes the input image in order to extract the header files and deployed hardware customer logic and used micro blaze processor over FPGA to perform SPIHT encoding and decoding. We have maintained the similar environment of implementation by considering the equivalent FPGA device (Spartan 3 XC400-5TQ144). The result shows that there is a considerable amount of improvement in all the performance parameters of proposed system in contrast to the recent work carried out by Mhamunkar [42].

The proposed system takes the odd and even components after it explores the high signal load in order to deploy it for generating low pass coefficients. This operation is exponentially speeded up by using enhanced Vedic multiplier, which has the further capability to minimize area as well as delay as compared to any traditional mechanism of radiological image compression. During the entire observation, it was found that proposed system has significantly lowered the dependencies on full adder as well as a half adder in comparison to usage of ripple carry adder. Implementation of DWT-based approach has further ensured retention of the maximum degree of signal quality of the reconstructed image.

VII. CONCLUSION

This research paper has presented an idea of performing compression of radiological images. Even after decades of research work towards image compression, this field has not witnessed a robust compression algorithm yet. In spite of availability of various compression schemes, the applicability of them in medical images are quite less owing to its dependencies on lossless compression scheme. After reviewing literature, it was explored that usage of DWT along with lifting schemes and FPGA are good possibilities to accomplish such lossless data during compression. Further, it was also found that potential characteristics of Vedic multipliers towards addressing the problems in compression are also left untapped in existing research work. Hence, a novel techniques have been discussed that jointly utilizes DWT, FPGA, and Vedic multiplier to accomplish an objective of cost-effective medical image compression. Our outcomes are also compared with similar kinds of schemes being recently published to find that proposed system exhibits a better balance between compression and image quality.

REFERENCES

[1] J. R. Haaga, D. Boll, Computed Tomography & Magnetic Resonance Imaging Of The Whole Body, Elsevier Health Sciences, 2016

[2] http://www.telemedproviders.com/telemedicine-articles/91-magnetic-resonance-imaging-mri.html

[3] H. Eren, J. G. Webster, Telemedicine and Electronic Medicine, CRC Press, 2015

[4] https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/ao2/report.html

[5] S. Avgousti, E. G. Christoforou, A. S. Panayides, S. Voskarides, C. Novales, L. Nouaille, C. S. Pattichis and P. Vieyres, “Medical telerobotic systems: current status and future trends”, BioMedical Engineering OnLine, 2016, DOI: 10.1186/s12938-016-0217-7

[6] M. J. Khan, A. Bhattacharyya, M. Alam, “Design Issues And Challenges Of Reliable And Secure Transmission Of Medical Images”, *eSAT Journals*, 2016, DOI: http://doi.org/10.15623/ijret.2014.0322034

[7] J. Ramsden, Bioinformatics: An Introduction, Springer, pp.34, 2015

[8] W. Kou, Digital Image Compression: Algorithms and Standards, Springer Science & Business Media, 2013

[9] M. El-Ghoboushi, Comparison of Lossless Image Compression Techniques based on Context Modeling, GRIN Verlag, 2015

[10] S. Gupta and Meenakshi, "A review and comprehensive comparison of image denoising techniques," *IEEE International Conference on Computing for Sustainable Global Development* (INDIACom), New Delhi, 2014, pp. 972-976, 2014

[11] D. Taubman, M. Marcellin, "JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards and Practice, Springer Science & Business Media, 2012

[12] P. V. Fleet, Discrete Wavelet Transformations: An Elementary Approach with Applications, John Wiley & Sons, 2011

[13] S-C Ou, H-Y Chung, W-T Sung, “Improving the compression and encryption of images using FPGA-based cryptosystems”, *Springer-Multimedia Tools and Applications*, vol.28, Iss.1, pp.5-22, 2006

[14] V. Dhandapani and S. Ramachandran, “Area and power efficient DCT architecture for image compression”, *Springer- EURASIP Journal on Advances in Signal Processing*, 2014

[15] Z. Fang, N. Xiong, L. T. Yang, X. Sun, and Y. Yang, “Interpolation-Based Direction-Adaptive Lifting DWT and Modified SPIHT for Image Compression in Multimedia Communications”, *IEEE Systems Journal*, Vol. 5, No. 4, December 2011

[16] P. Govindan, J. Saniie, “Processing algorithms for three-dimensional data compression of ultrasonic radio frequency signals”, *IEEE- IET Signal Processing*, Vol. 9, Iss. 3, pp. 267–276, 2015

[17] P. Govindan, B. Wang, P. Ravi, J. Saniie, “Hardware and software architectures for computationally efficient three-dimensional ultrasonic data compression”, *IEEE- IET Circuits, Devices & Systems*, pp.1-8, 2015

[18] M. H. Cabronero, I. Blanes, A. J. Pinho, M. W. Marcellin, J. S. Sagrista, “Analysis-Driven Lossy Compression of DNA Microarray Images”, *IEEE Transactions on Medical Imaging*, 2015

[19] M. R. Rehman, G. Raja and A. K. Khan, “Implementation of Lapped Biorthogonal Transform for JPEG-XR Image Coding”, *Intech Open Science*, 2012

[20] C. Yoon, W. Lee, J. H. Chang, T-K Song, and Y. Yoo, “An Efficient Pulse Compression Method of Chirp-Coded Excitation in Medical Ultrasound Imaging”, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 60, no. 10, October 2013

[21] A. Ahmad, N. H. Jaafar, A. Amira, “FPGA-based Implementation of 3-D Daubechies for Medical Image Compression”, *EMBS Conference on Biomedical Engineering and Sciences*, pp.683-688, 2012

[22] I. V. Anjaneyulu, P. R. Krishna, “FPGA Implementation of DWT-SPIHT Algorithm For Image Compression”, *International Journal Of Technology Enhancements And Emerging Engineering Research*, Vol 2, Iss.3, 2014

[23] G. Botella, C. García, and U. M. Base, “Hardware implementation of machine vision systems: image and video processing”, *Springer-EURASIP Journal on Advances in Signal Processing*, vol.152, 2013.

[24] Y. Li, W. Jia, B. Luan, Z-H Mao, H Zhang, M. Sun, “A FPGA Implementation of JPEG Baseline Encoder for Wearable Devices”, *41st Annual Northeast Biomedical Engineering Conference (NEBEC)*, pp.1-2, 2015

[25] M. Nagabushanam, C. P. Raj, S. Ramachandran, “Design and FPGA Implementation of Modified Distributive Arithmetic Based DWT – IDWT Processor for Image Compression”, *IEEE International Conference on Communications and Signal Processing*, pp.1-4, 2011

[26] Y-H Wu, Y-H Wu, L-X Jin, H-J Tao, “An improved fast parallel SPIHT algorithm and its FPGA implementation”, *2nd International Conference on Future Computer and Communication*, pp.191-195, 2010

[27] C. Xuesen, H. Liguo, D. Jinbo, “Design and implementation of FPGA-base diagnosis of medical image data acquisition equipment”, *IEEE-The Tenth International Conference on Electronic Measurement & Instruments*, pp.51-55, 2011

[28] Arish S, R.K.Sharma, “An efficient floating point multiplier design for high speed applications using Karatsuba algorithm and Urdhva-Tiryagbhyam algorithm”, *IEEE Global Conference on Communication Technologies*, pp.192-196, 2015

[29] R. K. Kodali, L. Boppana and S. S. Yenamachintala, “FPGA Implementation of Vedic Floating Point Multiplier”, *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems*, pp.1-4, 2015

[30] S.P.Pohokar, R.S.Sisal, K.M.Gaikwad, M.M.Patil, R. Borse, “Design and Implementation of 16 x 16 Multiplier Using Vedic Mathematics”, *International Conference on Industrial Instrumentation and Control*, 2015

[31] J. Sharma, C.Goyal, “High Speed 16- Bit Vedic Multiplier Using Modified Carry Select Adder”, *An International Journal of Engineering Sciences*, Vol. 17, 2016

[32] S. Singh, T. N. Sasamal, “Design of Vedic Multiplier using Adiabatic Logic”, *IEEE International Conference on Futuristic trend in Computational Analysis and Knowledge Management*, 2015

[33] G. Verma, S. Shekhar, K. S. Kant, V. Verma, H. Verma, and B Pandey, “SSTL IO Standard Based Low Power Arithmetic Design Using Calana Kalanabhyam On FPGA”, *International Journal of Control and Automation*, Vol. 9, No. 4, pp. 271-278, 2016

[34] A. E Vijayan, A. John, D. Sen, “Efficient Implementation of 8-bit Vedic Multipliers for Image Processing Application”, *IEEE International Conference on Contemporary Computing and Informatics*, pp.544-552, 2014

[35] D. Dennison and K. Ho, “Informatics Challenges—Lossy Compression in Medical Imaging”, *PMC-US National Library of Medicine, National Institutes of Health, Journal fo Digital Imaging*, vol.27, Iss.3, pp.287-291, 2014. 10.1007/s10278-014-9693-0

[36] S. Ravishkumar, V. Sridhar, “VCAR: Vedic Compression Algorithm Over Region of Interest on Radiological Image”, *IEEEInternational Conference on Emerging Research in Electronics, Computer Science and Technology*, 2015

[37] S. S. Kerur, P. Narchi, H. M Kittur, Girish V. A, “Implementation of Vedic Multiplier in Image Compression using DCT Algorithm”, *IEEE 2nd International Conference on Devices, Circuits and Systems*, 2014

[38] S. Ravishkumar, V Sridhar, “Computational Modelling of Image Coding using ROI based Medical Image Compression”, *International Journal of Computer Applications*, Vol.108, No. 5, December 2014

[39] S. Ravishkumar, V. Sridhar, “Design of Multiplier for Medical Image Compression Using Urdhava Tiryakbhyam Sutra”, *International Journal of Electrical and Computer Engineering* (Scopus-indexed), vol.6, No.3, 2016

[40] "Public Image Databases", Cornell University Vision and Image Analysis Group, Retrieved, 12 july, 2016, Link:- http://www.via.cornell.edu/databases/

[41] S. Todkar, P.V.S. Shastry, “Flipping Based High Performance Pipelined VLSI Architecture for 2-D Discrete Wavelet Transform”, *IEEE International Conference on Applied and Theoretical Computing and Communication Technology*, 2015

[42] N. S. Mhamunkar, B. S. Gayal, “Design And Implementation Of GENERIC 2-D Biorthogonal Discrete Wavelet Transform On FPGA”, *IEEE International Conference on Energy Systems and Applications*, 2015

Proposal of the Support Tool for After-Class Work based on the Online Threaded Bulletin Board

Kohei Otake
Faculty of Science and Engineering,
Chuo University
Tokyo, Japan

Yoshihisa Shinozawa
School of Science for Open and
Environmental Systems,
Keio University
Kanagawa, Japan

Tomofumi Uetake
School of Business Administration
Senshu University
Kanagawa, Japan

***Abstract*—In this paper, based on the assumption that after-class work in an exercise-based course accompanied by group work is done on an online threaded bulletin board system, the authors propose a support tool for the instructors. Specifically, while focusing on the factors that compose a discussion on the online bulletin board, the users who comment, the topics, and the items (keywords) to be discussed, the authors try to visualize the relationships among these factors as network diagrams. The authors also propose indexes, the comment degree and the activation degree, to evaluate communities formed there. Our experiments in which group work was actually implemented with the application of the proposed tool demonstrated that use of the network diagrams and the evaluation indexes served to distinguish the differences between those groups with properly-proceeding discussions and those without such discussions. The authors confirmed that this can enable the instructors to easily discover those students who do not participate in the discussion and groups with sluggish discussions.**

***Keywords*—Learning Support Tool; Online Threaded Bulletin Board; Network Analysis**

I. Introduction

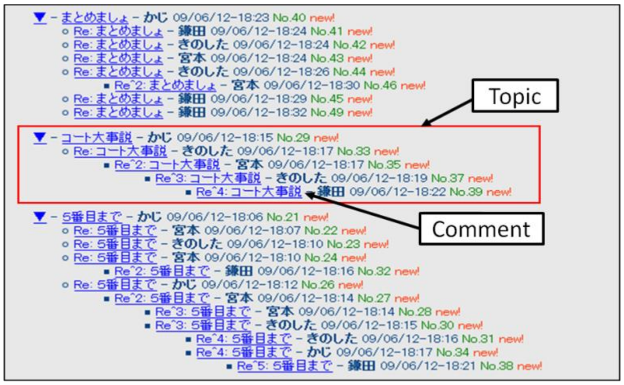
Many colleges and universities in Japan have recently focused on exercise-based courses accompanied by group work. Such courses have been implemented not only in introductory education typified by information literacy education, but also in specialized subjects including business exercises related to business administration. Introduction of this educational method accompanied by group work is believed to increase the depth of students' understanding. Additionally, it is expected to enhance student's communication skills and cooperativeness. [1] [2] [3].

This type of course usually handles one theme over several weeks. Therefore, the students are required to do their group work not only during the class hours, but also after the class without the presence of the instructors. However, the instructors have found it difficult to provide the students with necessary guidance for after class time. As a result, differences in the achievements of each student group might be produced.

Under these circumstances, a wide variety of educational support systems that assist students and the instructors have been proposed, such as e-mail, online bulletin board systems, and SNS tools[4][5]. For example, use of online bulletin boards offers not only easier information sharing within the group, but also enables the instructors to confirm the work progress of each group. Although these tools are utilized, the instructors still need to understand the work in progress of each group so that they can provide proper guidance. This creates even greater difficulty especially when the instructor has many groups of students to teach.

Based on the assumption that after-class work in an exercise-based course accompanied by group work is done on an online threaded bulletin board system (Fig. 1), in this paper, the authors propose a support tool that enables the instructors to easily understand the condition of progress of each group, while verifying the effectiveness of the proposed tool. Specifically, focus on the factors such as comment, the topics, and the items (keywords) to be discussed on the bulletin board; the authors try to visualize the relationships among these factors as network diagrams. Moreover, the authors propose indexes, the comment degree and the activation degree, to evaluate communities formed there.

The composition of this paper is as follows: Section 2 describes the current conditions of after-class work of exercise-based courses accompanied by group work, and the problems of such work that were identified and made clear by the previous studies conducted by the authors. Section 3 proposes the support tool for such work with the purpose of solving the problems described in section 2. Section 4 verifies the effectiveness of the proposed tool. Section 5 summarizes this paper and describes future issues.



▼ - まとめましょ - かじ 09/06/12-18:23 No.40 new!
- ○ Re: まとめましょ - 鎌田 09/06/12-18:24 No.41 new!
- ○ Re: まとめましょ - きのした 09/06/12-18:24 No.42 new!
- ○ Re: まとめましょ - 宮本 09/06/12-18:24 No.43 new!
- ○ Re: まとめましょ - きのした 09/06/12-18:26 No.44 new!
  - ■ Re^2: まとめましょ - 宮本 09/06/12-18:30 No.46 new!
- ○ Re: まとめましょ - 鎌田 09/06/12-18:29 No.45 new!
- ○ Re: まとめましょ - 鎌田 09/06/12-18:32 No.49 new!

▼ - コート大事説 - かじ 09/06/12-18:15 No.29 new!
- ○ Re: コート大事説 - きのした 09/06/12-18:17 No.33 new!
  - ■ Re^2: コート大事説 - 宮本 09/06/12-18:17 No.35 new!
    - ■ Re^3: コート大事説 - きのした 09/06/12-18:19 No.37 new!
      - ■ Re^4: コート大事説 - 鎌田 09/06/12-18:22 No.39 new!

▼ - 5番目まで - かじ 09/06/12-18:06 No.21 new!
- ○ Re: 5番目まで - 宮本 09/06/12-18:07 No.22 new!
- ○ Re: 5番目まで - きのした 09/06/12-18:10 No.23 new!
- ○ Re: 5番目まで - 宮本 09/06/12-18:10 No.24 new!
  - ■ Re^2: 5番目まで - 鎌田 09/06/12-18:16 No.32 new!
- ○ Re: 5番目まで - かじ 09/06/12-18:12 No.26 new!
  - ■ Re^2: 5番目まで - 宮本 09/06/12-18:14 No.27 new!
    - ■ Re^3: 5番目まで - 宮本 09/06/12-18:14 No.28 new!
    - ■ Re^3: 5番目まで - きのした 09/06/12-18:15 No.30 new!
      - ■ Re^4: 5番目まで - きのした 09/06/12-18:16 No.31 new!
      - ■ Re^4: 5番目まで - かじ 09/06/12-18:17 No.34 new!
        - ■ Re^5: 5番目まで - 鎌田 09/06/12-18:21 No.38 new!

Fig. 1. Example of online threaded bulletin board system

II. After-Class Work for Exercise-based Courses Accompanied by Group Work

A. Group work in exercise-based courses

Group work that is implemented during the exercise-based course includes discussions and debates. The purpose of group work stems from the following three points [6].

- Enhancement of the ability to analyze problems based on diversity of perspectives
- Enhancement of the ability to emerge with solutions and ideas
- Enhancement of the ability to cooperate while working in a group

Issues and themes, which are used in corporate training programs, are actually handled in many exercise-based courses. These issues and themes are based on educational materials, where the points to be discussed are clearly indicated and can be expressed with keywords. In many cases, therefore, the instructors find it easy to use them in their lectures.

B. Group work based on online threaded bulleting boards

As mentioned earlier, after-class work is essential when group work needs to be conducted over several weeks. For this reason, online bulletin boards and SNS tools are commonly used by which users can continue their work asynchronously and remotely. As described above, systems that support such work have also been proposed [4][5]. Actually, however, online threaded bulletin boards and similar systems are often used because they are easy to use and users can easily understand the flow of comments made [7]. Where group work is conducted on an online threaded bulletin board, on one theme, the users refer to and make comments on the same topic. This makes it easier to understand the condition of progress of each topic (Fig. 1).

C. Problems in after-class work

By using an online threaded bulletin board, the authors have conducted after-class work in exercise-based subjects for freshmen of private institutes and for those that are juniors and seniors of liberal arts universities in the Tokyo metropolitan area of Japan. The authors also analyzed the communities formed there [8][9][10]. Our analysis clarified that differences could be produced in the achievement level depending on the group, while disparities were also caused in the work level between the members within the group. Furthermore, when there were many groups, it became burdensome and difficult for the instructors to have a good understanding of the condition of each group.

As a result, these findings clarified that the instructors would need to check the checkpoints shown below for providing proper guidance in order for the students to do their work smoothly.

- Participation level of each group member
- Level of coverage of items to be discussed
- Properness of the proceedings (threads)
- Transition of the progress condition

III. Proposal of a Support Tool for After-Class Work

Based on the problems related to after-class work clarified in section 2, in this paper, the authors propose a support tool that makes it easier for the instructors to understand the condition of after-class work based on the online threaded bulletin board.

Specifically, the authors propose a function to visualize communities that are formed on the online threaded bulletin board and the indexes used for evaluating the communities formed.

A. Proposal of a function to visualize after-class work

In this paper, in order to achieve support for understanding the work condition on the online bulletin board, the authors focused on the important factors to understand the work condition, users who comment, topics, and items (keywords) to be discussed. Based on this focus, the authors propose a function that visualizes the four relationships shown below between two factors.

- Relationship between the users who comment and the topics
- Relationship between the users who comment and the items to be discussed
- Relationship between the topics and the items to be discussed
- Chronological changes in the above-mentioned relationships

The network-analysis [8] was used as the method for visualizing communities. Network analysis that the authors used here is a method that captures relationship patterns within the community as networks and that quantitatively expresses the patterns' structures as undirected graphs [11]. This method is used for analyzing communities [12][13][14][15][16][17].

In this paper, while expressing the relationships between two factors as one-on-one edges, the authors created undirected graphs based on these edges. In order to create these graphs, the node shape is changed according to each factor so that the node's attribute can be expressed depending on the node size.

In order to create network diagrams, the authors adopted the Kamada-Kawai method [18] based on the spring model, and used Pajek [19] which is software for network analysis. In the Kamada-Kawai method, strongly-related nodes are plotted in a near positional relationship; however, the absolute position of the node is not fixed. This means that a different network diagram is produced each time it is created. The purpose here is to understand the progress condition. Judging that the relative positional relationship between nodes would be important, the authors adopted this method. Fig. 2 shows the overview of the proposed visualization function.

1) Visualization of the relationship between the users who comment and the topics

Here, a network diagram is created by capturing the relationship between the users who comment and the topics as edges shown in Fig. 3. The size of the topic's node and that of the node of the user who comments are proportioned to the number of comments included within the topic and the number of comments made by the user.

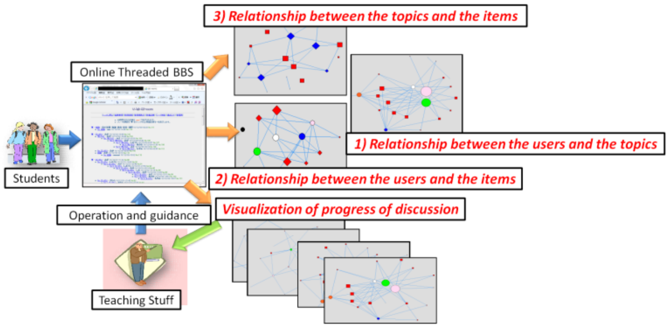

3) Relationship between the topics and the items
Online Threaded BBS
1) Relationship between the users and the topics
Students
2) Relationship between the users and the items
Operation and guidance
Visualization of progress of discussion
Teaching Stuff

Fig. 2. Overview of the proposed visualization function

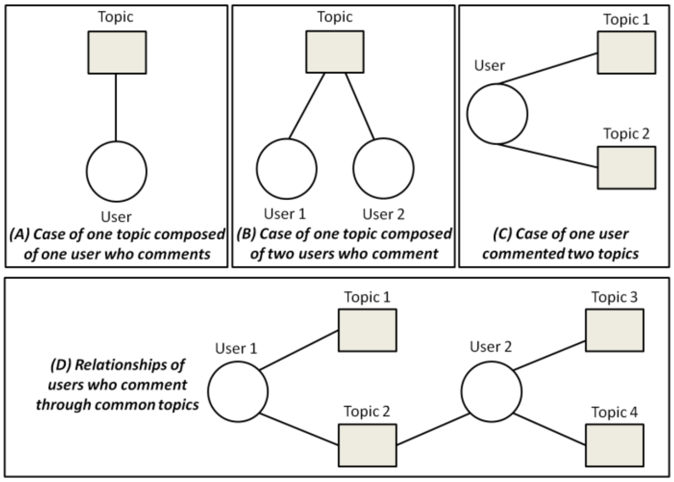

Topic
User
(A) Case of one topic composed of one user who comments
Topic
User 1
User 2
(B) Case of one topic composed of two users who comment
Topic 1
User
Topic 2
(C) Case of one user commented two topics
(D) Relationships of users who comment through common topics
User 1
Topic 1
Topic 2
User 2
Topic 3
Topic 4

Fig. 3. Relationship between the topic and the user who comments

The size of the node of the user who comments within this network diagram can make it easier to understand the participation level of each user who comments. The relationship between the topic and the user who comments can probably make it easier to understand the work progress condition.

2) Visualization of the relationship between the users who comment and the items to be discussed

The objective of work focused on in this paper is education as described in section 2.*A*. Therefore, the items to be discussed are clearly indicated, while keywords can be registered preliminarily. Here, a network diagram is created by using these keywords to connect the items (keywords) to be discussed with the users who comment as a one-on-one edge. As for keyword nodes, the node size expresses the number of reference counts.

The sizes of the nodes within this network diagram can probably make it easier to understand bias in the items to be discussed.

3) Visualization of the relationship between the topics and the items to be discussed

Here, the keywords included in the comments made and the topics referred to are connected with edges. As for the nodes of the topics and keywords, the node size similarly expresses the number of reference counts.

The relationship between the topics and the keywords in this network diagram probably makes it easier to understand the condition regarding the items to be discussed.

4) Visualization of chronolunnyouogical changes

Here, the network diagrams created in sections 3.*A*.1, 3.*A*.2, and 3.*A*.3 are saved in specified time intervals, while these diagrams are presented chronologically at the time of visualization. This function can serve to visualize the network formation process.

B. Proposal of indexes to evaluate communities

Next, indexes to evaluate communities that are formed on the online bulletin board are proposed. Based on previous studies, in this paper, the authors define the characteristics of ideal communities as the following two points.

- Many topics are commonly referred to by all the members
- All the users comment about the items (keywords) to be discussed, with less bias by the user

Here, the authors propose indexes which are the comment degree and the activity degree. The following shows the definitions and how to obtain these indexes.

Suppose that the number of the users who comment is I, and the number of the topics is J. The number of keywords stated by the user $i(i$=1,2,…, $I)$ in the topic $j(j$=1,2,…, $J)$ in a certain group is K_{ij}. The percentage of the keywords referred to by the user i in the topic j is expressed as R_{ij} (Equation (1)).

$$R_{ij} = \frac{K_{ij}}{\sum_{i=1}^{I} K_{ij}} \tag{1}$$

This percentage is obtained according to each topic by the user. The total value of all the topics is expressed as the comment degree H_i of the user i (Equation (2)). Namely, user with a higher comment degree H_i is determined to be making useful comments within the group.

$$H_i = \sum_{j=1}^{J} R_{ij} \tag{2}$$

This comment degree is then obtained according to each user in order to obtain the geometric average from all the comment degrees obtained. This geometric average obtained is regarded as the activation degree of discussion, S (Equation (3)), on the online bulletin board.

$$S = \sqrt[I]{\prod_{i=1}^{I} H_i} \tag{3}$$

This activation degree S becomes higher where many topics exist and where all the users are evenly discussing the keywords. Therefore, a group with a higher activation degree S is determined to be having discussions progressing properly.

Therefore, providing the instructors not only with network diagrams, but also with the comment degrees and activation degrees can make it easier to understand the condition of each user who comments and the discussion condition. This probably enables them to offer more proper guidance to the students.

IV. Evaluation

To verify the effectiveness of the proposed function and the evaluation indexes, the authors actually implemented a group work assignment on an online threaded bulletin board. Through this attempt, the authors analyzed the network diagrams, which were obtained from groups with high results and those with low results, and evaluation indexes.

A. Overview of evaluation experiments

The authors used a consensus game which is typically used for corporate training programs, such as the NASA game, as the group work assignment for experiments. The consensus game is an assignment where a certain theme is provided and the group members discuss the theme while trying to derive the conclusion within the specified time period. The purpose of this game is to gain a consensus from all the members of the group. In this evaluation experiment, the authors applied an assignment of ranking much-needed items that remain under a certain crisis situation. Targeting the juniors and the seniors of a private liberal arts university, the authors implemented the above-mentioned assignment on an online bulletin board. Since the purpose of this study is to support after-class work, the authors implemented two experiments consisting of a long-term experiment (a week) and a short-term experiment (about an hour) as a control experiment. A different theme was applied for each assignment.

- Experiment 1: To have the members come to a conclusion in a short period (about an hour)
- Experiment 2: To have the members come to a conclusion over a longer period (a week)

B. Experimental results

1) The result of the short-term assignment (Experiment 1)

Tab. 1 shows the results of experiment 1. The score of each group was calculated as the sum of squares of the difference in the correct answer and the answer of each group. Based on this score, groups were ranked from groups with lower scores.

TABLE I. The Result of the Short-Term Assignment

Group	**A**	**B**	**C**	**D**	**E**	**F**
Number of Users	4	4	4	5	4	4
Number of Topics	40	8	8	8	10	3
Number of Comments	111	104	71	110	51	53
Score	18	22	22	28	30	32
Rank	1	2	3	4	5	6
Activation Degree	6.01	1.48	1.24	0.91	1.62	0.23

TABLE II. The Result of the Long-Term Assignment

Group	**G**	**H**	**I**	**J**	**K**
Number of Users	5	5	6	5	5
Number of Topics	17	9	7	10	8
Number of Comments	102	45	56	52	67
Score	22	78	116	136	144
Rank	1	2	3	4	5
Activation Degree	1.79	0.81	0.64	0.86	0.00

2) The result of the long-term assignment (Experiment 2)

Tab. 2 shows the results of experiment 2. This table shows the results of implementing this assignment under an asynchronous and remote environment.

C. Evaluation of the function to visualize after-class work

1) Analysis based on the network diagrams of the relationships between the users who comment and the topics

Fig. 4 and Fig. 5 show the network diagrams of the relationships between the users who comment and the topics at the time point when experiment 1 and experiment 2 ended. Here, users who comment were shown circle, and topics were shown square.

In the network diagrams created by the Kamada-Kawai method adopted in this paper, the users who commented on the same topic or the topics consisted of the same users tended to be plotted in a near positional relationship. In the case of groups with high marks (group **A** in experiment 1 and group **G** in experiment 2), common topics discussed by all the users increased. In this network diagram, therefore, the nodes of the users who commented or the topic nodes tended to concentrate in the center. Experiment 2, which was conducted for a longer time outside of class hours, confirms that those users taking the lead in the discussion (core users) and free riders remarkably appeared, when compared to experiment 1.

Based on the summary of the results of Tab. 1 and Tab. 2, and Fig. 4 and Fig. 5, the following characteristics can be observed in the network diagrams of those higher groups.

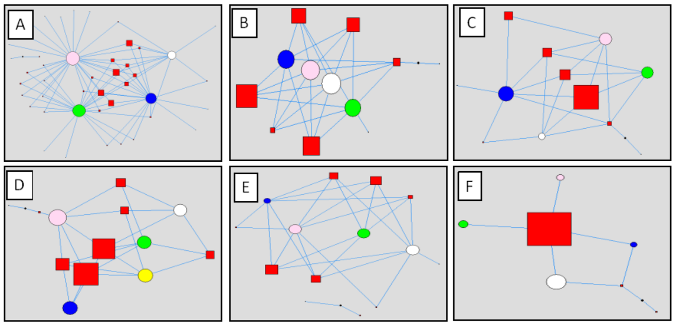


Fig. 4. The network diagrams of the relationships between the users who comment and the topics (Experiment 1)

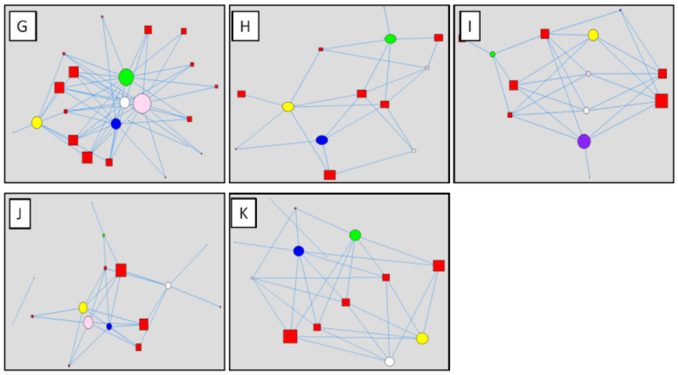


Fig. 5. The network diagrams of the relationships between the users who comment and the topics (Experiment 2)

- There exist many topics that are being discussed by all the members, and the nodes of the users who comment or the topic nodes tend to concentrate in the center
- All the members comment evenly, while no free riders exist

The above-described analysis shows that by using this network diagram, the instructors can provide more practical guidance by directing the students who make fewer comments to participate in the discussion, or by directing them to divide topics to develop the discussion when the topic nodes become too large. This can enable the instructors to offer support for the students with very little effort in order to promote well-balanced discussions.

2) Analysis based on the network diagrams of the users who comment and the items to be discussed

Next, Fig. 6 and Fig. 7 show the network diagrams of the users who comment and the items to be discussed at the time point when experiment 1 and experiment 2 ended. Here, users who comment were shown circle, and keywords were shown rhombus.

In experiments, keywords were defined as words that suggested the correct item names. As shown by Fig. 6 and Fig. 7, the network diagrams feature that the nodes of the users who made comments that concentrate in the center, and around these nodes keyword nodes are plotted. This is because when all the members participate in the discussion while commenting about all the keywords, the nodes of the users commenting concentrate in the center. The following characteristics can be observed in the network diagrams of those lower groups.

- Some users do not comment about the keywords
  - ➢ The nodes of these users are plotted away from the nodes of other users
- Some keywords are referred to less
  - ➢ The sizes of some keyword nodes are biased

These tendencies were more remarkably observed in experiment 2 which was conducted outside of class hours.

The above-described analysis shows that by using this network diagram, the instructors can provide more practical guidance by directing the students who less comment about the keywords. This also can make it easier when needed to indicate that some keywords being discussed are biased.

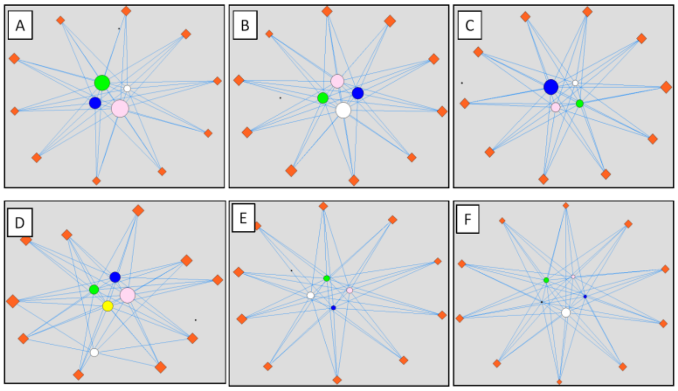


Fig. 6. The network diagrams of the users who comment and the items to be discussed (Experiment 1)

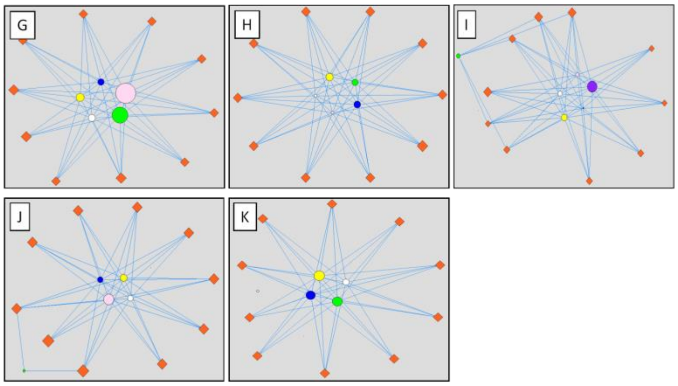


Fig. 7. The network diagrams of the users who comment and the items to be discussed (Experiment 2)

3) Analysis based on the network diagrams of the topics and the items to be discussed

Next, Fig. 8 and Fig. 9 show the network diagrams of the topics and the items to be discussed at the time point when experiment 1 and experiment 2 ended. Here, topics were shown square, and keywords were shown rhombus. From Fig. 8 and Fig. 9, the following characteristics can be observed in the network diagrams of those higher groups.

- Many topics are connected with all the keywords
  - ➢ The topic nodes concentrate in the center
- As for individual keywords, many topics are separately discussed
  - ➢ Topic nodes are plotted outside

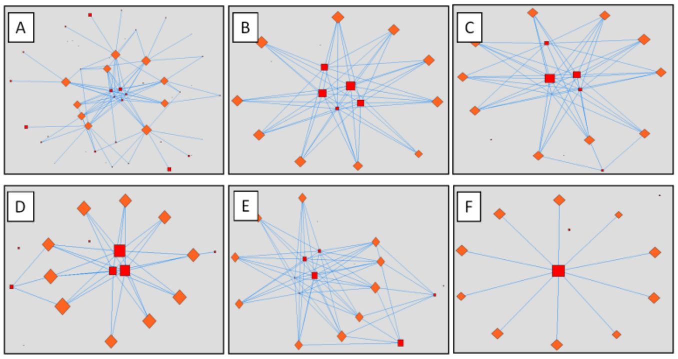


Fig. 8. The network diagrams of the topics and the items to be discussed (Experiment 1)

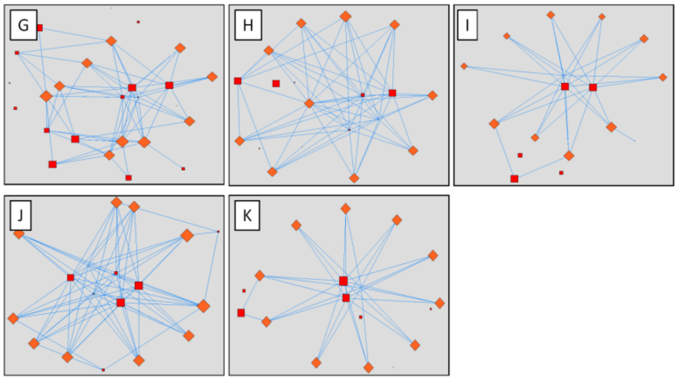


Fig. 9. The network diagrams of the topics and the items to be discussed (Experiment 2)

The above-describe analysis shows that use of this network diagram enables instructors to direct the students to set additional topics for the keywords where biased discussions are going on.

4) Analysis based on the network diagram with the discussion progress condition visualized

Fig. 10 and Fig. 11 show the network diagrams in which the daily progress condition of the highest group **G** and the lowest group **I** on the online bulletin board was visualized. The numbers in the upper left in these diagrams indicate the dates.

Comparison of Fig. 10 and Fig. 11 confirms that the discussion proceeds from the initial stage (the 3rd day) in the highest group **G**. Where the discussion does not proceed during the initial stage, therefore, it is important to encourage the users to advance their discussion. These network diagrams make it easier to find free riders. Therefore, use of these diagrams enables the instructors to guide those users with fewer comments to comment more during the initial discussion stage.

As described above, the authors were able to find that the proposed network diagrams served to distinguish the differences between groups with discussions proceeding properly and those without such discussions. As for those groups with sluggish discussions, therefore, use of this proposed function makes it easier to indicate the problems existing within the group.

Experiment 1 had only a short one hour implementation time. For this reason, experiment 1 had the tendency to have less bias, such as the number of comments that varied by each user, when compared to experiment 2. On the other hand, this problem related to bias could be observed remarkably in experiment 2 which was conducted outside of class hours. Therefore, in after-class work, use of the proposed tool for visualization probably makes it easier to understand the problems related to the discussion progress condition.

D. Analysis based on the evaluation indexes for communities

Tab. 1 and Tab. 2 show that the activation degree by the group is almost in the same order as the score-based evaluation. Use of the proposed indexes is probably determined to be valid in evaluating the discussion progress condition. However, as shown by group **E** in experiment 1 and by group **J** in experiment 2, there exist groups with high activation degrees but low scores. It is important to combine the indexes such as the number of topics or comments with network diagrams.

Next, Fig. 12 and Fig. 13 show the comment degree of the users who commented in each group. These diagrams confirm that the nodes of users with high comment degrees suggest the core users, while the nodes with low comment degrees suggest free riders. In addition, a number of users with high comment degrees exist in groups with high marks (groups **A** and **G**), while users with low comment degrees exist in groups with low marks (groups **J** and **K**).

The above-described analysis confirms that providing the instructors not only with network diagrams, but also with the indexes such as comment degrees and activation degrees can make it easier to understand the level of each user's participation in the discussion and the discussion progress condition of each group. This probably enables them to offer more proper guidance to the students.

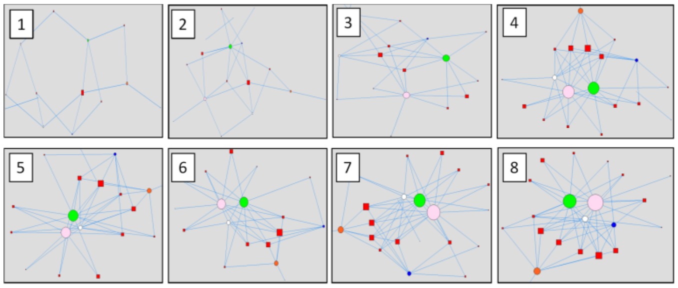


Fig. 10. Visualization of the discussion progress condition using network diagram (Group **G**)

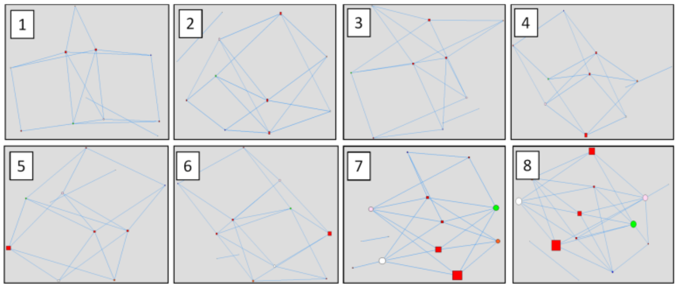


Fig. 11. Visualization of the discussion progress condition using network diagram (Group **I**)

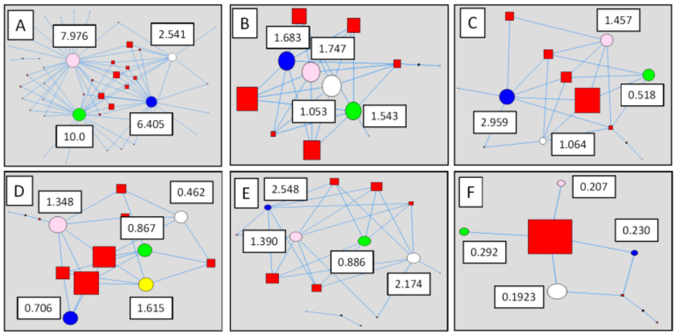


Fig. 12. Comment degree of the users who commented in experiment 1

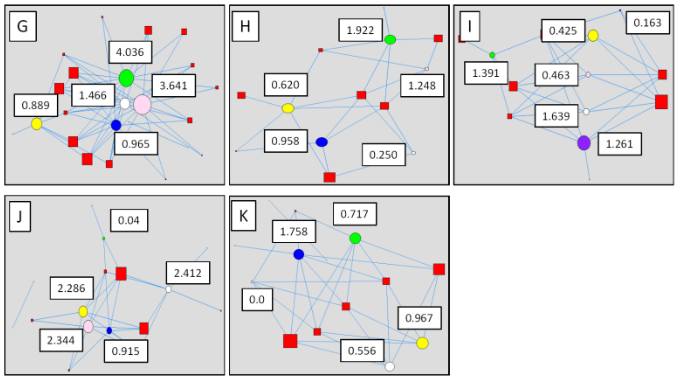


Fig. 13. Comment degree of the users who commented in experiment 2

Visualization based on the proposed support tool indicates the problems within the group without using the contents of comments made on the online bulletin board. In order to understand what kind of conversation is actually conducted within the group, it is necessary to view the contents of each comment made. The purpose of this support tool is to indicate the problems within the group. Therefore, the authors need to conduct more experiments by having the instructors use this support tool in the future. By doing so the authors can clarify to what degree the problems can be identified without viewing the contents of comments, and how much difference is observed when compared with the case of identifying the problems by checking all the comments that are made.

V. Conclusion and Future Issues

In this paper, based on the assumption that after-class work in an exercise-based course accompanied by group work is done on an online threaded bulletin board system, the authors proposed a support tool for the instructors.

Specifically, while focusing on the factors that compose a discussion on the online bulletin board, the users who comment, the topics, and the items (keywords) to be discussed, the authors tried to visualize the relationships among these factors as network diagrams. The authors also proposed indexes, the comment degree and the activation degree, to evaluate communities formed there.

Our experiments in which group work was actually implemented with the application of the proposed tool demonstrated that use of the network diagrams and the evaluation indexes served to distinguish the differences between those groups with properly-proceeding discussions and those without such discussions. The authors confirmed that this can enable the instructors to easily discover those students who do not participate in the discussion and groups with sluggish discussions.

In the future, in addition to examining more elaborate indexes, the authors are going to verify the further effectiveness of the proposed tool by conducting evaluation experiments with the instructors.

References

[1] Ichikawa. T and Nagata. M, “A Method for Group Learning in the Course of Management Information,”Journal of the Japan Society for Management Information, Vol. 12, No. 1, pp. 1-14 (2003)(in Japanese).

[2] Terakawa. K and Kawano. H, “Effect of Group Study for Information Literacy Education,” Proc. 66th National Convention of IPSJ, pp. 357-358 (2004) (in Japanese).

[3] Inoue. A, “Problem-Based Learning in Information Education, ” Journal of the educational application of infor-mation technologies, Vol. 8, No. 1, pp.41-45(2005) (In Japanese).

[4] Sawai. D and Miwa. J, “An Integrated Support System for Collaborative Learning in e-Learning,” IEICE Technical Report, ET2005-63, pp. 37-42 (2005) (in Japanese).

[5] Shimizu. Y, Nakajima. K, Komatsugawa. H, Kita. T and Yoshida. A, “Recent trends in e-learning based education system and technology,” Journal of the Institute of Electrical Engineers of Japan, Vol. 129, No. 9, pp. 596-615 (2009) (in Japanese).

[6] Uota. K, Ohsone. T, Ogiwara. S, Matsunaga. K and Miyanishi. Y, 「IT text, Basic Information Literacy」 Kyoritsu Shuppan, (2008) (in Japanese).

[7] Saito. M, “Collaborative Information Literacy Education Using an Online Discussion Board,” Research Reports of Yamawaki Gakuen Junior College, Issue. 42, pp. 32-43 (2004) (in Japanese).

[8] Shinozawa. Y and Uetake. T, “A Study of the BBS Communities which Assist Practice Classes by Using Network Analysis,” Journal of the Japan Society for Management Information, Vol. 15, No. 2, pp. 1-22 (2006) (in Japanese).

[9] Shinozawa. Y and Uetake. T, “Teaching support system for the group collaboration in the asynchronous learning environment,” European Conference on Computer-Supported Cooperative Work 2011, ECSCW 2011 Conference Supplement, pp.7-8 (2011).

[10] Uetake. T and Shinozawa. Y, “A Design of the Support System for the Group Collaboration to Cultivate Information Literacy Skills,” 13th International Conference on Human-Computer Interaction (DVD-ROM) (2009).

[11] Yasuda. Y, 「Practical network analysis - theories and techniques to solve relationships」, Shinyo-sha (2001) (in Japanese).

[12] Takahashi. M, Kitayama. S and Kaneko. I, “Measuring and Visualizing Organizational Awareness of Network Communities,” IPSJ Journal, Vol. 40, No. 11, pp. 3988-3999 (1999) (in Japanese).

[13] Fujita. K, Kamei. K, Jettmar. E, Yoshida. S,Kuwabara. K, “Network Analysis of a System Supporting the Formation of Cyber - communities,” IPSJ SIG Technical Reports, GW-39-1, pp. 1-6 (2001) (in Japanese).

[14] Mislove. A, Marcon. M., Gummadi, P.K. et al., “Measurement and analysis of online social networks,” IMC '07 Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 29-42 (2007).

[15] Yasutake. K,Tagawa. T,Yamakawa. O,Sumiya. T,Inoue. H, “An Analysis of Characteristic Properties of Communication Network Structure in E-Learning Courses,” Japan Journal of Educational Technology, Vol.31, No.3, pp.359-371(2007) (in Japanese).

[16] Wellman. B, Salaff. J, Dimitrova. D, Garton. L, Gulia. M and Haythornthwaite. C, “Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community,” Annual Reviews of Sociology, Vol. 22, pp. 213-238, (1996).

[17] Ebel. H, Mielsch. L and Bornholdt. S, “Scale-free Topology of E-mail Networks,” Physical Review E, Vol. 66, 035103(R), (2002).

[18] Kamada. T and Kawai. S, “An algorithm for drawing general undirected graphs information,” Processing Letters, Vol. 31, pp. 7-15 (1989).

[19] Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek/ (2017/01/24 author checked)

Impact Propagation of Human Errors on Software Requirements Volatility

Zahra Askarinejadamiri
Dept of software engineering and information system
Faculty of computer science and information technology
Universiti Putra Malaysia
Serdang, Selangor, Malaysia

Hazura Zulzallil
Dept of software engineering and information system
Faculty of computer science and information technology
Universiti Putra Malaysia
Serdang, Selangor, Malaysia

Abdul Azim Abd Ghani
Dept of software engineering and information system
Faculty of computer science and information technology
Universiti Putra Malaysia
Serdang, Selangor, Malaysia

Koh Tieng Wei
Dept of software engineering and information system
Faculty of computer science and information technology
Universiti Putra Malaysia
Serdang, Selangor, Malaysia

***Abstract*—Requirements volatility (RV) is one of the key risk sources in software development and maintenance projects because of the frequent changes made to the software. Human faults and errors are major factors contributing to requirement change in software development projects. As such, predicting requirements volatility is a challenge to risk management in the software area. Previous studies only focused on certain aspects of the human error in this area. This study specifically identifies and analyses the impact of human errors on requirements gathering and requirements volatility. It proposes a model based on responses to a survey questionnaire administered to 215 participants who have experience in software requirement gathering. Exploratory factor analysis (EFA) and structural equation modelling (SEM) were used to analyse the correlation of human errors and requirement volatility. The results of the analysis confirm the correlation between human errors and RV. The results show that human actions have a higher impact on RV compared to human perception. The study provides insights into software management to understand socio-technical aspects of requirements volatility in order to control risk management. Human actions and perceptions respectively are a root cause contributing to human errors that lead to RV.**

Keywords—Human factor; human errors; requirements volatility

I. Introduction

Software is developed based on the requirements of users which are obtained during the requirements gathering activity in the requirements engineering process in software development projects. The aim is to collect complete and unambiguous requirements. Nevertheless, not all projects are free from requirements changes or requirements volatility which involves additions, deletions, and modifications [1]. Frequent changes to requirements are a risk factor in software development projects [2].

Requirements engineering, which involves socio-technical aspects, is a critical and complex process. It has a vital role in reducing risks to a project and consequently increasing the success of software project [3, 4]. Among the elements to achieve success in software projects are technology, processes, and methods but the use of them is based on judgment and the decisions of human [5]. Thus, human aspects are among the main challenges in requirements engineering.

A variety of research and studies have addressed the technical aspects of requirements gathering and requirements volatility. They show the impact of on productivity [6], software defects [1] , and software release[7]. Moreover, not many studies focus on the factors that influence requirements volatility except that they are the communication between users and the developer and defined the methodology for requirements analysis and modelling [8]. A search on the ISI web of science shows that more than 70% of papers discuss the technical parts of software engineering and the software development process and less than 5% study the soft or human aspects of software development. Nevertheless, not many studies have focused on human factors as a vital component in controlling requirements volatility. Some researchers have explored human action and capability as reason of requirements volatility [9].

In this paper, we present a study on requirements volatility as a means to understand the impact of human errors on requirements gathering in requirements changes. It focuses on identifying and analysing human errors on requirements gathering which impact on requirements volatility. A model of human errors in relation to requirements volatility is proposed as a result of our extensive literature review. We employed quantitative approaches in validating the model. Thus, this study addresses the following research questions:

- Which human errors are relevant to requirements volatility in a software project?
- Which element of human errors has the most influence on requirements volatility?

Our results indicate the human errors which impact on software requirements changes are based on human action and goals, and human perceptions.

This paper is organized as follows. Section II provides the related work on human errors and requirements engineering. The methodology employed in this study is described in Section III, followed by the results and discussion in Section IV. Section V is the conclusion of the study.

II. Related Work

In the following section we derive and define the concept of human errors, human errors in RE and requirements volatility. We discuss around them to understand better relationship of human errors and requirements volatility.

A. Human Errors

The role of human is without doubt important to the successful development of software. For example, human reluctance to change may be important in controlling change in technically-based software processes or its tools [6]. However, in developing software we are often faced with development problems caused by human errors just like in other domain areas [10, 11]. In general, human errors are defined as any human activity which leads to not achieving the goals of the system [12, 13]. In software engineering perspective, not achieving the goals of the system means there are failures caused by faults originated from human errors. Thus, human errors are the root of the failures in a software project. Despite the occurrences of the failures, understanding of the nature of the failure in relation to human errors is important. For instance after the occurrence of a crash between Boeing 747 at Tenerife Island airport and the nuclear power plant accident, there was a concern to understand the nature of these disasters [14].

Human errors can occur in any phases of the software development lifecycle (SDLC). However, in this study, we focus on software requirements errors which occur in the first phase of the SDLC. During this phase, issues such as imprecise information, and incomplete or loss of data [15] cause failures or delays in projects. As most of these activities are affected by persons, human errors should be addressed and analyzed as a means to rectify such errors.

According to Helander et al. [16], human errors are of two types namely, phenomenological, which is concerned with error consequence, and those causing the error. The first category focuses on how the error occurred while the second group focuses on why. Omissions and substitutions are examples of the first group while slips, mistake, and cognitive errors fall into the second classification. In another view, the level of planning and intention can also lead to the emergence of various errors [17]. If the plan for the project is well designed but the activities involved do not effectively implement it, a slip emerges. Nevertheless, there are also cases where the action adheres properly to the plan but the plan itself is flawed. In both cases, human behaviour has a significant role in the occurrence of errors.

Some studies focus on human behaviour as the root cause of human errors which should be analysed based on human behaviour. They mention that human behaviour is based on knowledge, skills, and rules. Rule-based mistakes rely on the wrong rules or procedures. Knowledge-based behaviour emerges when there are no rules or procedures in a new environment. Mistakes here occur in situations of incomplete or wrong knowledge or interpretations. Skill-based errors occur when the wrong intention results in inappropriate execution of the plan [10, 18]. Overall, the occurrence of human errors is due to human behaviour which comprises rules, skills and knowledge.

A recent study on the assessment of human error on soft computing was conducted using fuzzy logic [19] which was used due to uncertainties in traditional human error risk assessments. Their model evaluated three risk factors in human error based on the fuzzy rule. Another study conducted a quantitative assessment of developer behaviour based on the data set [20]. They classified the behaviour of the developers in an automated way and applied statistical tools to analyse the model. In another work, software project human error reasons are classified into the attention of humans, communication error, and organization error (Harwood and Sanderson, 1986). The level of communication between stakeholders which is fully related to human personality is another issue in requirement engineering[21].

Apart from above, the psychological view of action slips are organized into 3 parts which are (a) errors in the formation of the intention (e.g., mode and description errors); (b) faulty activation of schemas (e.g., loss of intention and disordering of action components); and (c) faulty triggering (e.g., spoonerisms, blends, and intrusions of thoughts) [22]. According to human error theories, failure can happen in the goal, plan, and action stages of the human process [23]. Based on this idea, goals, and plans refer to something to be achieved and provide detailed steps for that purpose. The action is the implementation of the work to achieve a goal. Perception refers to the interpretation and evaluation of the action.

All the above mentioned studies focus on human errors in different dimensions. The following part focuses on studies on human error and requirement engineering which are at the root of many problems in requirements volatility.

B. Human Errors in RE

The various studies have been conducted identifying the relationship between human errors and requirements engineering. Lopes and Forster [23] focus on human error as one of the main reasons for RE failure. According to them, attention and memory error, communication error, organization errors, perception and interpretation errors, and violation are human errors that lead to RE failure. They presented a model for determining error types based on the type of problem although they also analyzed some aspects of human error. In addition, communication and interpretation errors have also been identified in requirements engineering and which are subject to user and developer communication in requirements gathering [24]. Usually, requirements elicited from customers are vague and incomplete and do not include adequate detailed information. Requirements are obtained through communication with stakeholders [25] and poor communication can reduce the quality of requirements gathering [26]. Undefined requirement process and misunderstanding are signs of poor communication in software requirements gathering [27].

Some techniques are presented for preventing defect in RE. In the paper titled “Preventing requirement defects: An experiment in process improvement”, the authors categorized software requirement defects as error source (where the true requirement has been ‘lost’); quality factor (functionality, usability, performance, etc.); related interface (user interface, third-party software, etc.); cost of handling and repair. Human and developer errors in this study are also considered as a source of defect in RE. Based on this study some techniques were presented which, if applied, can reduce failure of the project [28].

Scholars believe that human errors during the communication phase have a vital role to play in enhancing RE quality. They conducted their research using case studies. They classified communication and domain knowledge as two essential factors that impact on RE quality [29]. Simple omissions in communication can cause many challenges in requirements gathering [30]. In a similar study, researchers noted that trust relationship, increased knowledge, and better understanding are main elements that impact on the communication between users and developers. In one study, individual actors based on cognitive perspective, organizational factors, human flexibility, and human artfulness are the main contributors to human error that result in RE failure [31], while another study attributes it to organizational safety and human behaviour [32]. These factors are some aspects of human behaviour which cause human errors in the system.

In requirements engineering, verification and validation are important steps for development of the product as discussed in the paper titled “Challenges and practices in aligning requirements with verification and validation: a case study of six companies” [33]. The authors believe that weak communication is an example of weak RE that can cause many problems in software project such as invalid requirements, software quality problem, and wasted effort. Weak communication occurs through human errors, and the paper stressed the importance of human communication in requirement gathering in ensuring the success of a project.

Some studies have focused on the human personality and attitude on software engineering and apply theories such as Myers-Briggs Type Indicator (MBTI), the Big Five Personality Theory, and so on [5] [34][35].

All the above-mentioned studies focus on human errors and its metrics in requirements engineering while this research highlights requirements volatility in SDLC. Many scholars believe that requirements volatility is the root cause of project failure [1][8][36][37]. Due to the lack of studies on human errors and requirements volatility, this research focuses on this area especially in regard to the requirements gathering phase. Therefore, researchers try to list the human errors which impact on requirement changes based on the requirements engineering aspect. Scholars believe that requirements volatility is a metric of RE [8][38]. This study examines the human errors which impact on RV based on a review of RE papers.

C. Requirements Volatility

Based on the literature, requirements volatility is described as the following factors:

- Requirement instability: defined as requirements that fluctuate between the earlier and later stages, and differ at the start and end of the project [8][37][39].
- Requirement diversity: refers to the difficulty among shareholders in reaching agreement on the requirements and in customizing the software to one set of users requiring much effort to be expended in incorporating the requirements of the various users [8][37].
- Project Size: refers to the number of requirement changes including additions, modifications, and deletions in a software project [40]. Total development effort, project cost, and number of user representatives are involved [8][36].

III. Methodology

This section defines how the hypotheses were formed and validated, and describes the methodology and processes used to achieve the objectives of the research. The main research questions of this study are as follows:

- ➢ Which human errors are relevant to requirements volatility in a software project?
- ➢ Which element of human errors has the most influence on requirements volatility?

The first step in conducting the research based on the research questions is selecting research approach. It is a plan of research that determines the method of data collection and assumption validation, analysis, and interpretation [41].

A. Conceptual Framework

In order to present the hypotheses, researchers reviewed the papers to collect data for forming the model. By reviewing human errors on requirement engineering, we can collect human errors on requirement volatility.

TABLE I. CONSTRUCT AND ITEMS OF PRESENTED MODEL

construct	Items	Description
human action and goal	A1	Substitution of word or alphabet
	A2	Omitting word or sound
	A3	Gap in attention and memory failure
	A4	Omitting particular activity
	A5	Using or disregarding particular activity
	A6	Emotional makeup
	A7	Failure to set an objective
human perception	P1	Requirement gatherers' perception and interpretation
	P2	Cognitive behaviour
	P3	Understanding of requirement
Requirements volatility	RV1	Requirement fluctuate in earlier stage
	RV2	Requirement fluctuate in later stage
	RV3	Different in the requirement of start and final in the project
	RV4	Difficulty for stockholders to reach agreement on requirements
	RV5	Difficulty to customize the software to one set of users
	RV6	A lot of effort had to be spent in incorporating the requirement of the various user
	RV7	Number of requirement change include add, modified and delete in a software project
	RV8	Total development effort, project cost and number of user representatives involved

Thus, the selected human errors for this research are described in two categories as follows:

- Goal and Action: is defined as requirements gathering for software development to be achieved by the requirement gatherer based on his plan and action. The human errors in this case are failure to set an objective, substitution of word or alphabet, omitting word or sound, gaps in attention and memory failure, omitting a particular activity, and using or disregarding a particular activity.
- Perception: relates to the act of perceiving, interpreting, and evaluating the results of the requirement gathering action. The activities in this type of human error are requirement gatherers' perception and interpretation, cognitive behaviour, and understanding of requirements.

Although ideally the requirements for software projects should be complete and unambiguous before the design phase, in real-life situations changes to them are unavoidable. Requirements volatility leads to redesigning, recoding, and retesting and may even result in the failure of the project [42]. There is a direct relationship between requirements volatility and defect density [1]. Understanding human errors is a key element for managing requirements volatility in order to achieve success in a software project.

Therefore, the relationships were analysed based on the above mentioned elements on requirements volatility and human errors. In this study, RV is considered a dependent variable and human errors as independent variables.

To control RV, it is necessary to manage or minimize human error. Based on the literature reviews, human errors are considered as significant elements in this research. Human errors are classified as human actions, goals, and perceptions.

The hypothesized model of this paper is presented in Fig 1. It shows the correlation between human errors and requirement volatility. Also all constructs of the presented model is described in Table I.

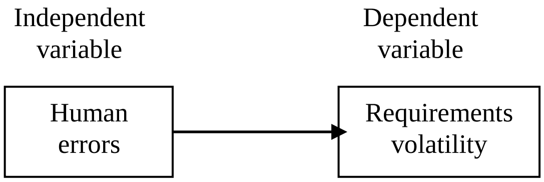


Fig. 1. Hypothesized model of this study

B. Phase2: Model Evaluation

Following the presentation of the model a quantitative approach was done to test and validate it using data that was collected and analysed. One of the main aims of quantitative research is to understand the relationship between variables [43]. This research is conducted based on the Structural Equation Modelling (SEM) technique with the aim of achieving a convergence of opinions concerning human errors and requirement volatility from persons who have experience in software requirement gathering. Due to the lack of theories in this research, this study is exploratory and a questionnaire was administered among participants to gather data for analysing human errors and software requirement volatility.

C. Participants

A sample population of software requirements gatherers should preferably be chosen as representatives in this study. Unfortunately, there is no data available for such a population. As this study also faced financial and time constraints in selecting a sample, persons with experience in software requirements gathering were selected as respondents.

D. Sample Size

Sample size has a significant role in statistical analysis and in this study it is based on the statistical analysis technique that will be used for the research. In this research SEM and SPSS is used for data analysis. There is no consensus on the exact sample size for SEM and researchers have different ideas on that. SEM needs an appropriate sample size in order for the estimation to be reliable and valid. Some scholars mention that a sample size of 200 is a critical number for analysing structural equation modelling [44][45] while Kline [46] suggests a number between 200 to 400. In general a minimum sample size of 200 is appropriate. This study involved 215 respondents.

E. Data Collection

This section describes the procedure for the validation hypothesises of this research. In order to examine them, the survey approach was done. The analyses of human errors on requirement volatility are based on responses to the questionnaire which were distributed online and by hand. Online questionnaire in google doc and *kwicksurvey* were developed to facilitate respondents. The online questionnaires shared in the social media were those related to requirement engineering and software engineering as shown in Table II. Apart from online data collection, due to the accessibility of researchers in Iran and Malaysia, the questionnaires were distributed in Technology Park Malaysia and Cyberjaya where most Malaysian software companies are located.

The questionnaire was designed in order to understand the relationship between requirements volatility and human errors. It was based on the literature reviews in order to provide the aims of this research. A 5-point Likert scale with a range of strongly disagrees to strongly agree was used.

TABLE II. ONLINE QUESTIONNAIRE SHARED IN SOCIAL MEDIA

Social media	Group
LinkedIn group	Requirement engineering, requirement engineering specialist group (RESG),Requirement management and analysis, Mobile_software_developer, Mobile software development ,Software developer engineer in Test (SDET), Software designer and development, Software development management professional, Swedish association for requirement engineering(SARE), software developer, software and technology, IT and software project management, computer and software engineering professionals groups
Facebook group	Software engineer, I am a software engineer, Software engineering, Software developer
Yahoo group	Developers_for_ever, Leandevelopment, Requirement-engineering

F. Data Analysis

SPSS version 21 was used for the statistical analysis. This study addressed treatment of the missing data, tested for normality of data, and identified outlier by using SPSS. An analysis was also made of the demographic profiles of respondents for the study. Additionally, Exploratory Factor Analysis (EFA) was done by SPSS to summarize the variable in a different group and analyse the information. It classified the factors of the research model based on principal component analysis (PCA) by the Varimax rotation method. EFA was employed based on common factor model to summarize variables for factors [47]. In addition, the reliability of the construct and model was checked by Cronbach alpha test with a value greater than 0.7 applied to confirm the reliability of the model.

SEM is an accumulation of statistical methods that look for clarifying connections among different variables. It empowers analysts to look at the interrelationships among different dependent and independent variables [47]. The basis for selecting SEM for investigation in this research is its capacity to test relationships of complicated models having multivariate variables. Further, it offers excellent statistical procedures for managing complex models [47] as well as flexibility in statistical tests for the measurement of invariance [48]. Confirmatory Factor Analysis (CFA) allows for the analysis of relationships between dependent and independent variables (measurement model) [49]. SEM consists of two step which are measurement modelling and structural modelling. In order to perform SEM, CFA test will be done. It identifies the relationship between constructs and indicators which will be done by CFA using the AMOS software.

In order to do measurement modelling evaluation, Confirmatory Factor Analysis (CFA) in AMOS was used to examine the relationship between variables and relationships between the constructs and indicators. Model checking based on goodness of fit was conducted and the hypothesised model was improved the fit. This research used structural modelling to test the interrelationship between dependent and independent variables hypothesized in this research. Structural modelling was done to test the correlation between human error and requirements volatility and how human errors impact on RV.

IV. RESULT AND DISCUSSION

The majority of respondents in this study were male and had more than 5 years of work experience in software development projects. In this sample, we addressed a variety of respondents from different organizations, countries, and positions. The demographic details of respondents are shown in Table III.

This section discusses construct validity and reliability and presents the results of the exploratory factor analysis (EFA) and the structural equation modelling (SEM). First, the EFA tests are used to identify the relationship between measured variables. After identifying the relationship, the model fit was tested by CFA. In order to conduct EFA and CFA tests, the data should be normal with no missing data and outliers. Data screening were done to check for missing data, normality and outlier. In this study there were six missing values which were replaced by using the median technique which is a good means to address low levels of missing data. Also, normality of data was checked by analysing Skewness and Kurtosis and the results show they were between -2 and +2 which shows normality of data. Outliers are defines as an observation that are distinctively different from other values [47], and problematic ones should be identified in research. The two main outliers are univariate and multivariate outliers with the former referring to data consisting of extreme values on variables while the latter is a combination of unusual values [46]. Also, identifying outliers will be discussed in the section on SEM prior to conducting the CFA.

TABLE III. DEMOGRAPHIC DETAILS OF RESPONDENTS ((N=215)

Variable	Category	frequency	%
Gender	Male	130	61
	Female	82	38.5
	Others	3	0.5
Age	21-30	92	42.8
	31-40	100	46.5
	Over40	23	10.7
Job title	Software developer	76	35.3
	Software engineer	59	27.4
	System analysts	9	4.2
	Function analyst	5	2.3
	Business Analyst	13	6.0
	Information architect	3	1.4
	Others	50	23.3
Type of organization	Governmental	2	.9
	Semi-governmental	8	3.7
	Private	205	95.3
Work Experience	Less than 5 years	74	34.4
	5-10 years	94	43.8
	More than 10 years	45	20.9

A. Construct validity and reliability

Based on above criteria, the results show that Human Action and Goal and Human Perception and RV are valid constructs. The construct validity of the instrument used in the research has been assessed through convergent validity and discriminant validity. Convergent validity alludes to examining whether the degree of relationship between two measures of construct in theory is valid in fact. Average variance extracted (AVE) and construct reliability (CR) were used to calculate convergent validity. In order to assess convergent validity, the cut-off AVE point should be greater than 0.5 and CR should be greater than AVE. The results of this study are shown in Table IV indicating that the value of CR and AVE are more than the cut-off point. Discriminant Validity refers to whether a construct is truly distinct from others. It is assessed using Maximum Shared Value (MSV), AVE, and Average Shared Square Variance (ASV). MSV should be less than AVE and ASV should be less than AVE to establish that the construct's discriminant validity is an accepted criteria [47].

TABLE IV. RESULT OF MEASUREMENT MODEL VALIDITY

	CR	AVE	MSV	ASV
Action	0.886	0.526	0.504	0.291
perception	0.784	0.548	0.504	0.252
Requirement volatility	0.895	0.517	.0784	0.039

Based on results in Table IV, it can be said that the values support discriminant validity. Reliability is another important issue which should be tested. For this test, this study employed SPSS to provide the Cronbach alpha. The results show the alpha= .884 for 18 items of this study which is greater than the cut-off point and shows that the hypothesized model is reliable.

B. Exploratory factor analysis (EFA)

Before conducting EFA, the outliers should be identified. In this study, there was no univariate outlier, because the Likert Scale was used for responses and participant cannot respond beyond this 5-scale range. In order to identify the multivariate outlier Mahalanobis distance (D^2) test was used to measure the distance between each observation and compared to the mean of the observations [50]. However, the results show that there were a few outliers in this study. Hair et al. [47] mention that the removal of outliers can improve the multivariate analysis but has the risk of decreasing generalizability. Also they believe that in a sample size greater than 100, if $MD^2/2$*number of items measures do not exceed 3 or 4, the cases remain in the data and are not considered as an outlier.[47] In this study the levels for the $MD^2/2$*number of items of suspected outliers are less than 3 which do not exceed a critical value. Thus, in this research there was no evidence of multivariate outlier and all data remains in the research.

EFA is a statistical method for identifying the structure of relative variables through extraction and rotation. Extraction is used to determine the factors of the variables while rotation is used to provide a pattern for better interpretation [47]. To conduct the EFA for this study, principal components analysis (PCA) was used for the extraction and Varimax rotation was performed for the rotation by SPSS. Table V shows the results of KMO and Bartlett's. The KMO value is used to measure sample adequacy and suitability of data for construction which in this research is 0.914 indicating that the number of data for analysis is acceptable and suitable. Communality is a criterion of the EFA which is extracted using PCA and shows the common factor analysis. The extraction values in the communality table VI indicate the proportion of each variable's variance that can be explained by the principal components. Items with higher values are well represented in the common factor space while variables with low values are not. In this study the communalities of items vary from .557 to .682. The lowest communality value is the Failure to set an appropriate objective.

TABLE V. KMO AND BARTLETT'S TEST

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.910
Bartlett's Test of Sphericity	Approx. Chi-Square	1808.68
	Df	153
	Sig.	.000

One of the important questions in this study is identifying the factors based on items and this was done using principal component analysis. Table XII is extracted based on Eigen values greater than one. Variances of 25.81 %, 23.28%, and 11.909% are explained by the first, second, and third factors respectively while the remaining variance is explained by the other four factors. In order to achieve a clear pattern of loading, rotational strategies were conducted to identify the factors. Varimax rotation was selected for this study to maximize the variance on the new axes. The factors were extracted using the Rotated Component Matrix. By performing EFA, the number of constructs and indicators were extracted. Table VII shows the results of constructs and indicators. It can be said that the results confirm the existence of the three factors based on the selected items. Based on these results, the correlations of items were extracted and three constructs of this study namely, Goal and Action, Perception, and Requirement Volatility were identified.

TABLE VI. COMMUNALITY OF ITEMS

Items	Extraction
Substitution of word	.652
Omitting Word	.580
Gap in attention	.578
Omitting activity	.554
Disregarding particular rule	.626
Emotional make up	.683
Failure to set appropriate objective	.560
Perception and interpretation	.714
Cognitive behaviour	.679
Understanding of requirement	.658
Requirement Fluctuated in earlier stage	.570
Requirement fluctuated in later stage	.597
Different requirement identified	.589
Difficult stockholders to reach agreement	.631
Effort had to be spent in incorporating	.594
Difficult to customize software	.561
Number of requirement change	.555
Total development effort	.600

TABLE VII. ROTATED COMPONENT MATRIX

	Component		
	1	2	3
Difficult stakeholders to reach agreement	.791		
Total development effort	.773		
Requirement fluctuated in later stage	.759		
Different requirement identified	.757		
Requirement Fluctuated in earlier stage	.749		
Effort had to be spent in incorporating	.735		
Difficult to customize software	.732		
Number of requirement change	.727		
Disregarding particular rule		.806	
Emotional make up		.781	
Substitution of word		.743	
Omitting Word		.716	
Failure to set appropriate objective		.711	
Gap in attention		.710	
Omitting activity		.661	
Perception and interpretation			.784
Understanding of requiremnt			.766
Cognitive behaviour			.737

One of the important questions in this study is on identifying the factors based on items. Principal component analysis was conducted to extract the factors. Based on Eigen values greater than one Table VII is extracted. Variances of 25.81%, 23.17%, and 11.80% are explained by the first, second, and third factors, respectively while the remaining variance is explained by the other four factors.

C. Structural Equation Modeling (SEM)

In SEM, there are three basic types of model fit indices namely, absolute, incremental, and parsimonious [51]. Their criteria are presented in Table XIII. In this research, three factors were considered in the measurement model, which are Human Action and Plan (A), Human Perception (P), and Requirement Volatility (RV). These factors are measured by using ten items. The measurement model of this research is evaluated by testing the maximum likelihood (ML) as provided by AMOS. The initial results of CFA in this research are shown in Table XIII. The chi-square statistics (χ2=156.881,df=1.188) was significant at p<0.05 and reveal that the fit of data to this measurement model should be accepted. It shows that the presented model, 95% can generalize to real model. Due to the sensitivity of the chi-square statistic to the sample size and its normality it is not appropriate to rely only on this item. Therefore, other fit indices such as AGFI, CFI, RMR, RMSEA, PCLOSE, PCFI, and IFI are used to assess the measurement model. In order to reflect model fit, Jaccard and Wan [52] recommend reporting at least three fit tests comprising one absolute, one relative, and one parsimonious. The results of these criteria are presented in Table XIII.

The results of this study show that the value of RMR=0.018, CFI=0.989, RMSEA= 0.026, IFI= 0.989, AGFI=0.911, PCFI=0.853, and Pclose=1 and indicate that the value of the model fit is above the cut-off point and it can be said that the model is fit. We can refine the model in order to achieve a better model fit [46] and some techniques are presented for that purpose. The standardize loading factor should be greater than 0.5 to be acceptable in model [53]. Additionally, a standard residual value between 2.58 and -2.58 is acceptable [47]. An assessment of the results shows that the values of the standard residual and loading factors are above the cut-off point value and are acceptable. It can say that the model is fit. Thus the Final measurement model is presented in fig 2.

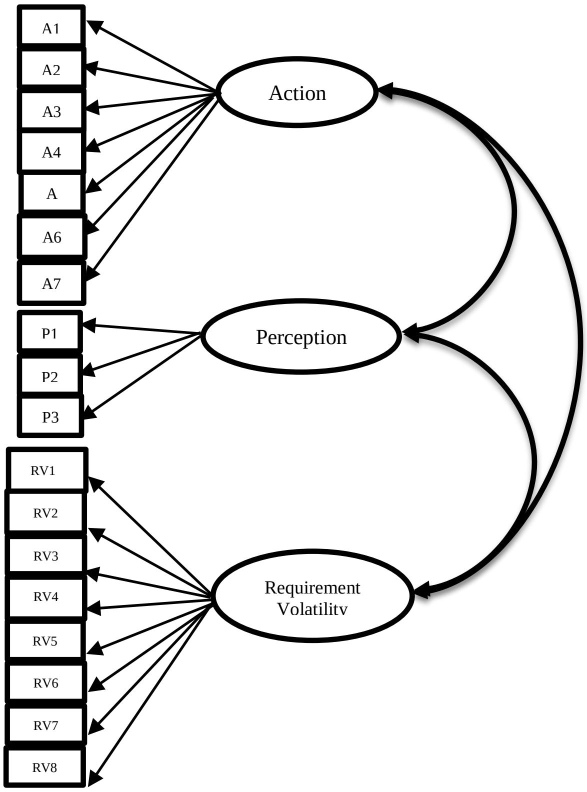


Fig. 2. Final Measurement Model

D. Structural Model Evaluation and Hypotheses Testing

This part discusses hypotheses testing. The two hypotheses of presented model in this research are presented in Table VIII. According to H1, there is a positive relationship between human goal and action and Requirement Volatility. Similarly, H2 indicates the positive correlation between human perception and requirement volatility. This section tests the relationship between these independent variables with requirement volatility. The final model was drawn by AMOS (Fig 3).

TABLE VIII. STRUCTURAL MODEL EVALUATION AND HYPOTHESES TESTING

Construct	code	hypothesise	**Hypothesised Relationships(positive/negative)**
Human Action	A	H1	A → RV
Human perception	P	H2	P → RV

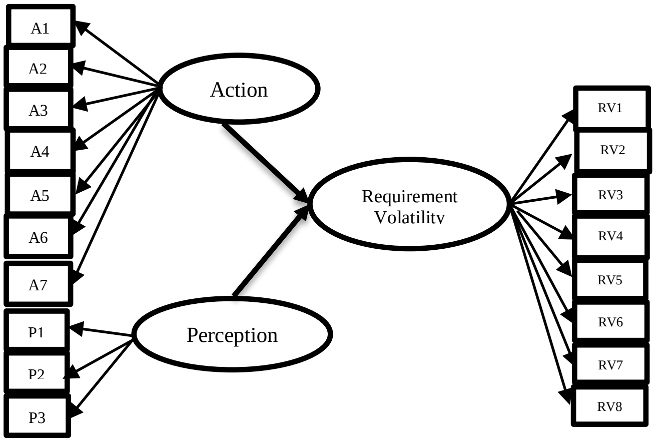


Fig. 3. Structural Model

A coefficient parameter assessment should be done for evaluating this model. This model has two latent constructs which are defined by ten items. In order to evaluate the model, factor covariance which is the critical ratio will be checked to be greater than 1.96 for an estimate. In this case, it can be said that the factor covariance is significant. Consequently the coefficient value is less than 0.05 and is statistically significant [47]. The critical ratio is calculated by dividing the regression weight (estimate) by the standard error (SE). The results of the coefficient parameter assessment for the two factors are presented in Table IX. As shown, the assessment for human action and perception was at the significant level p ≤.05. Additionally standardize regression weight of human action and goal, and human perception estimated 0.51 and 0.27 which indicate that human errors based on action and goal have a 51% impact on RV compared to 27% for the impact of perception.

TABLE IX. REGRESSION WEIGHT

	Estimate	S.E.	C.R.	P
RV <--- A	0.51	.13	3.934	.001
RV <--- P	-0.27	.074	-3.677	***[1]

TABLE X. HYPOTHESIS TESTING

Construct	code	hypothesise	**Hypothesised Relationships(positive/negative)**	**Supported**
Human Action	A	H1	A → RV	Yes
Human perception	P	H2	P → RV	Yes

The results of the tests reveal that hypotheses H1was positive and H2 was negatively statistically significant. The results suggest that standardized estimates for these hypotheses (β = 0.51, 0.27, respectively) indicate statistical significance and thus show support for these hypotheses. These results show the statistically significant connection between human errors and requirements volatility. People with higher human error can increase the requirements volatility and it is similar with the findings reported in studies such as Lopes et al. [23], Decker [27] ,and Andrew and Brad [54].

In addition to this general finding, a more detailed analysis of the results of this study indicated the following. Table XI shows the Standardized Regression Weights of the indicators of IV constructs. Based on these results, it can be stated that A1 (Substitution of word or alphabet) and A6 (Emotional makeup) have the most influence on the goal and action constructs of human errors and consequently on RV. In contrast, A7 (Failure to set an objective) has less impact on goal and action constructs of human errors and consequently on RV. The results show that the root of RV are based on human errors are goal and action of human at work. Thus, it is necessary to decrease RV by controlling human goal and action. This is significant information for software manager to improve their requirement gatherer skill in goal and action skill of them for requirements gathering to decrease human errors and consequently requirements volatility. In the human errors perception construct, P1 (Requirement gatherers' perception and interpretation) and P3 (Understanding of requirements) respectively have the highest and lowest impacts on this construct and on RV subsequently.

TABLE XI. STANDARDIZED REGRESSION WEIGHTS OF INDICATORS OF IV CONSTRUCTS

Indicators--->construct	Estimate
A1--->A	.778
A2--->A	.721
A3--->A	.720
A4--->A	.709
A5--->A	.702
A6--->A	.785
A7--->A	.660
P1--->P	.736
P2--->P	.709
P3--->P	.687

E. Implication for SE practice and research

Based on theoretical implication, this study proposes a model in the context of human errors in software development. This study answers the call to examine the causes of RV in requirements gathering. The findings suggest that human errors impact on RV in the requirements gathering. This research has provided extended knowledge in the domain of RV from a developers' perspective. Additionally, this research attempts to reduce the paucity of research on the role of human errors on RV. Unfortunately, very little study is known about the human errors on RV in software requirements gathering. Another significant contribution of this study is the instrument used for collecting the research data. There is a dearth of instruments for measuring RV and human errors constructs. This instrument or questionnaire has been carefully designed, developed, and statistically validated and thus can be used for future research particularly in the area of RV and human error. Previous researches have focused on the technical causes of RV while this study highlights the socio-technical aspects. In short, in investigating the root causes of RV this study has focused on human errors, especially in communication for requirements gathering.

Findings of this research study have practical implication for managers of software companies. First, in order to control requirements volatility in software development activities, project managers must have a good understanding of how to

[1] Less than 0.001

facilitate and cultivate effective communication between requirements gatherers and users. Emphasis should be given into identifying and understanding the enablers and impediments towards communication for software requirements gathering. From our findings, serious consideration in the areas of human errors must be taken in order to manage the communication errors issues. Further, the emphasis given will assist project managers or team leaders improve the capability and behaviour of requirements gatherers in communication during the software requirements gathering stage. We believe that our effort fills the gap due to lack of understanding and prescription on the socio-technical aspects of RV in software development. We propose a model of human errors on RV that shows key human errors that have the potential in stimulating RV, and directly impacting on the quality of gathered requirements. In addition, this study signifies that human errors were identified as the elements that impact on RV with human action and goal having the most impact on it while human perception are other human errors which impact on RV. Hence, software managers should consider the human errors of the requirements gatherer as a means to manage and achieve better RV.

V. Future Work

This paper developed an integrated human errors model that provides a systematic way to understand RV due to human factors comprising human action and human perception. Several beneficial areas for future research, however, remain to be explored. For example, the results of current research are limited to RV and future research may apply or replicate this study in other software development domains. Also, there are some other human factors that impact on RV and could be apply to this model as future work.

VI. Conclusion

This work provides the position of human errors to the processes of RE and, consequently, RV in order to improve them by minimizing errors. Thus, first of all, the importance of different causes of human error in software requirements gathering was collected based on qualitative research. Then these hypotheses and models were validated by analysing the collected response of participants. By identifying the different root causes of human errors, we confirmed that they have an impact on requirements volatility in software requirements gathering. Software managers are frequently confronted with the risk of requirements changes which give rise to many issues in their maintenance management operations. Knowing the roots of this challenge enables them to be controlled more effectively. In short, some human errors based on the constructs of goal and action and perception which impact on RV were presented. The result shows that goal and action of humans has a higher impact on RV compared to their perceptions.

References

[1] Y. K. Malaiya and J. Denton, "Requirements volatility and defect density," in Software Reliability Engineering, 1999. Proceedings. 10th International Symposium on, 1999, pp. 285-294.

[2] Q. Wang, D. Pfahl, and D. Raffo, "Making Globally Distributed Software Development a Success Story," in International Conference on Software Process, ICSP 2008, Leipzig, Germany, May 10-11, 2008.

[3] N. Juristo, A. M. Moreno, and A. Silva, "Is the European industry moving toward solving requirements engineering problems?," IEEE software, vol. 19, pp. 70-77, 2002.

[4] D. Mishra, A. Mishra, and A. Yazici, "Successful requirement elicitation by combining requirement engineering techniques," in Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the, 2008, pp. 258-263.

[5] R. Feldt, L. Angelis, R. Torkar, and M. Samuelsson, "Links between the personalities, views and attitudes of software engineers," Information and Software Technology, vol. 52, pp. 611-624, 2010.

[6] M. Lane and A. Cavaye, "Management of Requirements Volatility Enhances Software Development Productivity," in Australian Conference on Requirements Engineering (ACRE), Geelong, Australia, 1998.

[7] G. Stark, A. Skillicorn, and R. Ameele, "An examination of the effects of requirements changes on software releases," CROSSTALK The Journal of Defence Software Engineering, pp. 11-16, 1998.

[8] D. Zowghi and N. Nurmuliani, "A study of the impact of requirements volatility on software project performance," in Software Engineering Conference, 2002. Ninth Asia-Pacific, 2002, pp. 3-11.

[9] N. Ibrahim, W. M. W. Kadir, and S. Deris, "Propagating Requirement Change into Software High Level Designs towards Resilient Software Evolution," in 2009 16th Asia-Pacific Software Engineering Conference, 2009, pp. 347-354.

[10] D. A. Norman, The design of everyday things: Revised and expanded edition: Basic books, 2013.

[11] J. Rasmussen, "Human errors. A taxonomy for describing human malfunction in industrial installations," Journal of occupational accidents, vol. 4, pp. 311-333, 1982.

[12] G. S. Walia and J. C. Carver, "A systematic literature review to identify and classify software requirement errors," Information and Software Technology, vol. 51, pp. 1087-1109, 2009.

[13] J. Senders and N. Moray, "Human error," Cause, prediction and reduction, Orono, Univ. Maine, 1991.

[14] J. Reason, Human error: Cambridge university press, 1990.

[15] D. Firesmith, "Common Requirements Problems, Their Negative Consequences, and the Industry Best Practices to Help Solve Them," Journal of Object Technology, vol. 6, pp. 17-33, 2007.

[16] M. G. Helander, T. K. Landauer, and P. V. Prabhu, Handbook of human-computer interaction: Elsevier, 1997.

[17] K. Harwood and P. Sanderson, "Skills, rules and knowledge: A discussion of Rasmussen's classification," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1986, pp. 1002-1006.

[18] D. Embrey, "Understanding human behaviour and error," Human Reliability Associates, vol. 1, pp. 1-10, 2005.

[19] P.-c. Li, G.-h. Chen, L.-c. Dai, and Z. Li, "Fuzzy logic-based approach for identifying the risk importance of human error," Safety science, vol. 48, pp. 902-913, 2010.

[20] E. Di Bella, A. Sillitti, and G. Succi, "A multivariate classification of open source developers," Information Sciences, vol. 221, pp. 72-83, 2013.

[21] P. Lenberg, R. Feldt, and L. G. Wallgren, "Behavioral software engineering: A definition and systematic literature review," Journal of Systems and Software, vol. 107, pp. 15-37, 2015.

[22] D. A. Norman, "Categorization of action slips," Psychological review, vol. 88, p. 1, 1981.

[23] M. E. R. F. Lopes and C. H. Q. Forster, "Application of human error theories for the process improvement of requirements engineering," Information Sciences, vol. 250, pp. 142-161, 2013.

[24] R. G. Mays, C. L. Jones, G. J. Holloway, and D. P. Studinski, "Experiences with defect prevention," IBM Systems Journal, vol. 29, pp. 4-32, 1990.

[25] D. Würfel, R. Lutz, and S. Diehl, "Grounded requirements engineering: An approach to use case driven requirements engineering," Journal of Systems and Software, vol. 117, pp. 645-657, 2016.

[26] Endres and H. D. Rombach, A handbook of software and systems engineering: Empirical observations, laws, and theories: Pearson Education, 2003.

[27] S. W. Dekker, "Illusions of explanation: A critical essay on error classification," The International Journal of Aviation Psychology, vol. 13, pp. 95-106, 2003.

[28] S. Lauesen and O. Vinter, "Preventing requirement defects: An experiment in process improvement," Requirements Engineering, vol. 6, pp. 37-50, 2001.

[29] Marnewick, J.-H. Pretorius, and L. Pretorius, "A perspective on human factors contributing to quality requirements: A cross-case analysis," in IEEE Industrial Engineering and Engineering Management, Singapore, 2011, pp. 389-393.

[30] K. M. de Oliveira, F. Zlot, A. R. Rocha, G. H. Travassos, C. Galotta, and C. S. de Menezes, "Domain-oriented software development environment," Journal of Systems and Software, vol. 72, pp. 145-161, 2004.

[31] S. Viller, J. Bowers, and T. Rodden, "Human factors in requirements engineering:: A survey of human sciences literature relevant to the improvement of dependable systems development processes," Interacting with Computers, vol. 11, pp. 665-698, 1999.

[32] M. A. Teruel, E. Navarro, V. López-Jaquero, F. Montero, and P. González, "An empirical evaluation of requirement engineering techniques for collaborative systems," in Evaluation & Assessment in Software Engineering (EASE 2011), 15th Annual Conference on, 2011, pp. 114-123.

[33] E. Bjarnason, P. Runeson, M. Borg, M. Unterkalmsteiner, E. Engström, B. Regnell, et al., "Challenges and practices in aligning requirements with verification and validation: a case study of six companies," Empirical Software Engineering, vol. 19, pp. 1809-1855, 2014.

[34] J. E. Hannay, E. Arisholm, H. Engvik, and D. I. Sjøberg, "Effects of personality on pair programming," Software Engineering, IEEE Transactions on, vol. 36, pp. 61-80, 2010.

[35] P. Holtkamp, J. P. Jokinen, and J. M. Pawlowski, "Soft competency requirements in requirements engineering, software design, implementation, and testing," Journal of Systems and Software, vol. 101, pp. 136-146, 2015.

[36] S. Ferreira, J. Collofello, D. Shunk, and G. Mackulak, "Understanding the effects of requirements volatility in software engineering by using analytical modeling and software process simulation," Journal of Systems and Software, vol. 82, pp. 1568-1577, 2009.

[37] R. Govindaraju, A. Bramagara, L. Gondodiwiryo, and T. Simatupang, "Requirement Volatility, Standardization and Knowledge Integration in Software Projects: An Empirical Analysis on Outsourced IS Development Projects," Journal of ICT Research and Applications, vol. 9, pp. 68-87, 2015.

[38] M. Davis, N. Otero, K. Dautenhahn, C. L. Nehaniv, and S. D. Powell, "Creating a software to promote understanding about narrative in children with autism: Reflecting on the design of feedback and opportunities to reason," in 2007 IEEE 6th International Conference on Development and Learning, 2007, pp. 64-69.

[39] T. Javed and Q. S. Durrani, "A study to investigate the impact of requirements instability on software defects," ACM SIGSOFT Software Engineering Notes, vol. 29, pp. 1-7, 2004.

[40] P. Mohagheghi and R. Conradi, "An empirical study of software change: origin, acceptance rate, and functionality vs. quality attributes," in Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on, 2004, pp. 7-16.

[41] D. Crowther and G. Lancaster, Research methods: Routledge, 2012.

[42] M. Bano, S. Imtiaz, N. Ikram, M. Niazi, and M. Usman, "Causes of requirement change-a systematic literature review," in Evaluation & Assessment in Software Engineering (EASE 2012), 16th International Conference on, 2012, pp. 22-31.

[43] F. Gravetter and L.-A. Forzano, Research methods for the behavioral sciences: Cengage Learning, 2011.

[44] H. a. Schaubroeck, "Confirmatory modelling in organizational behaviour/human resource management: issues and applications," Journal of Management, 1990.

[45] K. A. Markus, "Principles and Practice of Structural Equation Modeling by Rex B. Kline," Structural Equation Modeling: A Multidisciplinary Journal, vol. 19, pp. 509-512, 2012.

[46] R. B. Kline, Principles and practice of structural equation modeling: Guilford publications, 2015.

[47] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, Multivariate data analysis vol. 6: Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[48] P. J. Swerdzewski, Should we worry about the way we measure worry over time? A longitudinal analysis of student worry during the first two years of college: ProQuest, 2008.

[49] D. Harrington, Confirmatory factor analysis: Oxford University Press, USA, 2008.

[50] M. Byrne, Structural equation modeling with AMOS: Basic concepts, applications, and programming: Routledge, 2013.

[51] Hooper, J. Coughlan, and M. Mullen, "Structural equation modelling: Guidelines for determining model fit," Articles, p. 2, 2008.

[52] J. Jaccard and C. K. Wan, LISREL approaches to interaction effects in multiple regression: Sage, 1996.

[53] V. Kachitvichyanukul, K. Sethanan, and P. Golinska-Dawson, Toward Sustainable Operations of Supply Chain and Logistics Systems: Springer, 2015.

[54] J. Ko and B. A. Myers, "A framework and methodology for studying the causes of software errors in programming systems," Journal of Visual Languages & Computing, vol. 16, pp. 41-84, 2005.

TABLE XII. TOTAL NUMBER OF FACTORS AND VARIANCE EXTRACTED

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	% of Cumulative	Total	% of Variance	Cumulative %	Total	% of Variance	% of Cumulative
1	34.098	34.098	34.098	34.098	25.817	25.817	4.648	25.821	25.821
2	21.311	55.410	21.311	55.410	23.171	48.988	4.190	23.28	49.101
3	5.379	5.379	60.789	11.801	60.789	5.379	2.144	11.909	61.01

TABLE XIII. GOODNESS OF FIT STATISTICS OF CFA MODEL

			Absolute fit			Incremental fit measures		Parsimony fit measure		
	χ2	DF	χ2/df	RMSEA	RMR	CFI	IFI	PCLOSE	AGFI	PCFI
criteria			1<X2/df<3	<0.05	small RMR~ good fit RMR=0: exact fit	≥0.90	≥0.9	> 0.5 :good fit	>0.8	>0.5
Result of this study	150.765	132	1.162	0.026	0.018	.989	.989	1	0.911	0.853
Note:	**χ2**= Chi-square; df = degree of freedom; RMSEA = Root mean square error of approximation; RMR= Root Mean Square Residual ; CFI = Comparative fit index; IFI= Incremental Fit Index; AGFI = Adjusted goodness of fit index ; PCFI= Parsimony-adjusted Comparative Fit Index									

Helpful Statistics in Recognizing Basic Arabic Phonemes

Mohamed O.M. Khelifa
TES Research Team
ENSIAS School of Engineering
Mohammed V University in RABAT
Rabat, Morocco

Yahya O.M. ElHadj
Doha Institute, Doha, Qatar
SAMoVA Research Team, IRIT
Paul Sabatier University
Toulouse, France

Yousfi Abdellah
FSJES-souissi
Mohammed V University in RABAT
Rabat, Morocco

Mostafa Belkasmi
TES Research Team
ENSIAS School of Engineering
Mohammed V University in RABAT
Rabat, Morocco

***Abstract*—The recognition of continuous speech is one of the main challenges in the building of automatic speech recognition (ASR) systems, especially when it comes to phonetically complex languages such as Arabic. An ASR system seems to be actually in a blocked alley. Nearly all solutions follow the same general model. The previous research focused on enhancing its performance by incorporating supplementary features. This paper is part of ongoing research efforts aimed at developing a high-performance Arabic speech recognition system for learning and teaching purposes. It investigates a statistical analysis of certain distinctive features of the basic Arabic phonemes which seems helpful in enhancing the performance of a baseline HMM-based ASR system. The statistics are collected using a particular Arabic speech database, which involves ten different male speakers and more than eight hours of speech which covers all Arabic phonemes. In HMM modeling framework, the statistics provided are helpful in establishing the appropriate number of HMM states for each phoneme and they can also be utilized as an initial condition for the EM estimation procedure, which generally, accelerates the estimation process and, thus, improves the performance of the system. The obtained findings are presented and possible applications of automatic speech recognition and speaker identification systems are also suggested.**

Keywords*—automatic speech recognition (ASR); speech recognizer; phonemes recognition; speech database; hidden Markova models (HMMs)

I. Introduction

The most communal way for humans to communicate is through sounds made during speech operation. Thoughts and ideas are exchanged via speech. One person speaks and the other receives the message by means of their ears. Automatic speech recognition (ASR) is the process by which a computer is capable of recognizing and acting upon spoken language or utterances using particular algorithms [1-5]. It is a branch of artificial intelligence (AI) and is related to various areas of knowledge, including informatics, linguistics, acoustics, and pattern recognition. An ordinary ASR system consists of a microphone unit, speech recognition engine, computer, and a certain form of audio/visual/action output. The Applications of an ASR system can be classified into two main areas. One is dictation, and the other is human-computer dialogue applications. In the dictation area, the broadcast news dictation technology has been incorporated into information extraction and retrieval technology, and many application systems such as retrieval systems and automatic voice document indexing. In the human-computer interaction area, a variety of experimental systems for information retrieval through spoken dialogue were investigated. A common ASR application is the automated conversion of speech into written text, which has the capability to increase output effectiveness and enhance access to diverse computer applications such as word processing, email, remote control, using phones, language identification, speaker identification, and archiving and language acquisition.

By using speech as input, ASR applications reduces the more traditional manual input techniques via keyboards and mousses, making it helpful as an alternative input technique for people with disabilities. ASR performance may be affected by various factors, including the quality of the inputted speech, the technology design, the surrounding environment and speaker characteristics.

In spite of the remarkable advances in signal processing, computational architectures, algorithms and hardware, ASR systems is still a topic of an active research and ideal systems are still far from reached [6]. Thus, the most important research issues should be attacked in order to advance to the ultimate goal of fluent speech recognition.

In speech recognition, it is uncomplicated to recognize isolated words but the main challenge is to recognize continuous speech. There are two parts for any ASR system: the language model and the acoustic model. The language model indicates the status of word sequences to be recognized: are they common or rare? Thereby, the acoustic model is used to model the sounds we produce when we speak. For a small vocabulary, it’s easy to model the acoustics of individual words. As vocabulary size grows, it becomes impractical to record sufficient spoken examples of all words and so we need to model acoustics at a lower level. The state-of-the-art ASR systems do not rely on the whole words in both training and decoding process due to the enormous quantity of words that may exist in a speech corpus in addition to the necessity to have sufficient spoken examples for each word. Contrariwise, a successful ASR system uses smaller parts of words or sub-word units of words that are commonly designed by phoneticians or expert in linguistics. This set of sub-word units is referred to as phonemes.

Most of the current successful ASR systems are based on hidden Markov models (HMM) in which each phoneme is modeled by a set of HMM states. A 3 emitting states with left-to-right HMM topology are commonly used for each phoneme independent of its length. Thus, the question that arises is whether this number of states is sufficient for certain phonemes or is it greater or fewer than what is needed? One of the main matters in ASR system is to determine the number of HMM states that reflects the correct length of each phoneme occurrence in a speech corpus.

Despite the sizable utilization of speech recognition technologies in foreign languages likes English and French, Arabic the rarity of mature ASR-based applications, especially for language teaching and learning. One renowned application of Arabic Speech Recognition is the teaching of Classical Arabic (CA) sound system. Although classical Arabic is not utilized in everyday communication, it is required for learning the Holy Quran (The Muslim Holy Book) and the old Arabic poetry heritage. Moreover, it can open the door for various sorts of Islamic applications.

The present paper is part of ongoing research efforts aiming to develop a high-performance Arabic speech recognition system for learning and teaching purposes. First stages of these efforts were dedicated to the development of particular Arabic speech database including ten different speakers and more than eight hours of speech collected from recitations of the Holy Quran in which all Arabic phonemes are included. Speech signals of this speech database were manually and accurately segmented and labeled on three levels: word, phoneme, and allophone. Next, two baselines HMM-based recognizers were built to validate the speech segmentation on both phoneme and allophone levels and also to examine the intended recognition accuracy in both recognizers.

This current stage investigates a statistical analysis of certain distinctive features in Arabic phonemes in order to incorporate them later into the speech recognition process for the aim of improving the performance of our baseline HMM-based recognizers. The distinctive features which have been investigated in this work are phoneme durations, mean durations of phonemes, median of the duration for each basic phoneme, median of the durations, frequency and probability occurrences for each basic phoneme. Analysis and interpretations were performed to determine which of these distinctive features can significantly enhance systems performance. In HMM modeling framework, the statistics provided can be helpful in establishing the appropriate number of HMM states for each phoneme which generally increases the speed and recognition accuracy. The phonemes statistics can also be utilized as an initial condition for the Expectation-Maximization estimation procedure and hence accelerates the estimation process, or it can be utilized as a wanted model itself. Also, the probability of the neighboring two phoneme clusters is helpful information which is not yet integrated in the adjustment of speech characteristics of possible words from a dictionary.

The rest of the article is organized as follows: section 2 summarizes our research efforts accomplished towards the ultimate goal. Section 3 describes the motivation of the presented work. Section 4 introduces a brief overview of the previously developed speech database. In Section 5 we present the methodology used for statistics extraction. Section 6 gives the details of the statistical analysis implemented. Finally we conclude the paper by giving a conclusion in section 7.

II. Research Efforts Summary

As findings of a previously funded research project [7], two baseline HMM-based systems for phonemes and allophones [8, 9] were constructed using the mentioned speech database. The number of allophones in the speech database is 110 plus a silence unit which is counted as normal allophone indicating short pauses during the recitations, while the number of phonemes is 60, which represents almost half of the number of allophones. All speech units were modeled by an HMM with three emitting states for both levels to capture their acoustic properties. And for each state, a Gaussian Mixture Models (GMMs) were also associated to designate the characteristics of the sound portion at this state. The Mel-frequency cepstral coefficients (MFCCs) were used as cepstral acoustical features. For each Hamming window of 10 ms, a vector of 39 MFCCs was extracted. These coefficients are the first twelve MFCC plus their first and second derivatives to capture the sound's static features at this portion. Also, the energy plus its first and second derivatives were appended to identify the sound's dynamic features at the same portion. The hidden Markov model toolkit (HTK) was employed to train and test the HMMs for both systems. The word error rates (WERs) obtained for these recognizers were respectively 8% and 12% for phonemes and allophones.

Our current efforts focalized on the development of an elaborate system, by firstly considering the basic sounds and then looking for their distinctive features to determine which ones will be particularly helpful to well identify their phonological variation. To this end, we have adopted the speech database to be annotated in terms of basic phonemes. We mean by the basic phonemes the basic sounds without any phonological variation and even without considering the sounds gemination (the doubling). They are 32 phonemes. Their list and their associated codes are shown in the table 2.

The new version of the speech database was utilized in all efforts yet accomplished, including an HMM-based recognizer for basic Arabic sounds [10], an enhanced Arabic phonemes recognizer using duration modeling techniques [11] and an accurate HSMM-based system for Arabic phonemes recognition [12]. In the last implemented system for the basic Arabic phonemes [12], the average recognition rates obtained are about 99 %.

III. Background and Motivation

Automatic Speech recognition (ASR) seems to be actually in a blocked alley. Nearly all solutions are of the same general model [13]. The research focused on enhancing its performance by integrating supplementary elements. Such an approach yielded better results but it must be admitted that there is a limit which cannot be overrun without modification of the general scheme. The method based on hidden Markov models (HMMs) with features of fixed frames length has found its utility in numerous applications. However, it does not seem to be effective enough to transcribe properly any spoken language with a large vocabulary. There are several reasons. Some of them are very straightforward in their nature. The dictionary-based ASR system will never work correctly for out-of-dictionary words. Grammar models will not deal correctly with incorrectly spoken utterances while humans very often can.

ASR system tries to recognize speech via these matching techniques, while humans can easily understand it and adopt it to mistakes and unusual words. This causes the mentioned limit of the classical ASR approaches. The standard ASR approach is, indeed, based on guess and luck in few steps of its procedures. The inputted speech is segmented into frames without any motivated rules. HMM attempts to find the closest transcription on the basis of speech features which, indeed, a kind of guessing. Such approach works well enough for plainly spoken words with a limited vocabulary. Noise, the speaking rate and the large vocabulary cause many exclusions and data missing which HMM cannot deal with correctly. Another major problem is that people do not speak as carefully as they write, while we anticipate a transcription produced by an ASR system to be of the grade of our typed texts.

It has also to be admitted by both ordinary users and researchers, that when we speak we do not, at all times, follow grammar rules and, furthermore, the mistakes in pronunciation involve various exceptions independently of the dictionary size used. This is why adopting a hypothesis using related language rules and a limited dictionary does not always work satisfactorily. The same issues take place in the case of names, out-of-language words, and the mispronounced phonemes, etc. ASR system attempts to adopt the inputted speech to the language rules and the static vocabulary, which, in certain cases, leads to supplementary distortions and hence to degradation in system performance.

There is no straightforward solution for the above-described problems. In this work, we suggest the use of collected phoneme statistics in a target language in order to be used as, for instance, a support for the dictionary if there is a difficulty in associating matching features to one of the words to be recognized in the vocabulary.

The most outstanding research works carried out on continuous speech is based on statistical approaches specifically Hidden Markov Models (HMM). Many HMM-based ASR systems for continuous Arabic speech have reached various levels of recognition accuracy and encouraging performances which have been achieved [14-18]. The accuracy of recognition is usually measured by the correct percentage of recognized phonemes. The HMM-based ASR systems performance is affected by various factors including the existence of noise; the number of HMM states associated with each phoneme; the phoneme combination used and the phonemes length. Enhancing performance of the present ASR techniques needs the examination of these cited factors in order to localize and recognize the regions of enhancement.

Nonetheless, no fully statistical analysis at the phoneme level has been implemented on this speech database of classical Arabic sounds used in this work. Statistical analysis of Arabic phonemes gives a comprehensible vision of phonemes behavior and provides the capability to regulate this behavior by investigating the gathered statistics. For example, the frequency of a specific phoneme in a speech database can be employed to correct its misrecognition during the decoding process. This means replacing this misrecognized phoneme by the highest probably one.

Furthermore, the average duration of a particular phoneme can also be utilized to estimate the number of HMM states that are most appropriate for recognizing it. Additional statistical information such as mode (the midst value in a set of values) and median (the most frequent value in a set of values) are advantageous in addressing the misrecognized phonemes during the decoding process. In this paper, we present a full statistical analysis of Arabic phonemes which can be employed for the purpose of enhancing performance of our baseline HMM-based systems by reducing the word error rate (WER) factor.

IV. Speech Database of Sounds

The Arabic language is the official language of about 300 million speakers around the world. It is the religious language of all Muslims around the world, regardless of their native language. It is the official language in all Arab countries and the 6th most widely utilized language in terms of first language speakers. Arabic can be categorized into two main variants: Classical Arabic (CA) and Modern Standard Arabic (MSA). CA is an old literary form of Arabic, which is the most formal type and is the language of the Holy Quran and the old Arabic poetry. MSA is the current standard form of Arabic, which is utilized in official communications in Arabic countries, broadcast news, formal speeches, etc. Although there is no big difference between today's Arabic (MSA) and that spoken by the early Arabs (CA), due to the fact that Arabic is one of the most stable languages throughout history, yet there are some idiosyncrasies as to the way of pronunciation.

One of the main barriers faced by the development of ASR applications for Arabic speech is the rarity of suitable sound databases commonly required for training and testing statistical models. This problem is seriously approached when dealing with classical Arabic language since most of the corpora available nowadays are specifically oriented towards what is known as Modern Standard Arabic (MSA) and its sub-forms (i.e. dialects). To remedy this problem and to assist the development of ASR applications for classical Arabic language, a speech database covering all classical Arabic sounds was designed on the basis of Quranic recitations. The speech corpus was developed in a previously funded project by Al-Imam Muhammad ibn Saud Islamic University in Saudi Arabia with the support of King Abed Al-Aziz City for Science and Technology (KACST). Because of the difficulty of developing this kind of corpora, only a part of the Holy Quran was regarded. Recitations of ten male speakers were recorded in an appropriate environment under the supervision of an expert of the holy Quran pronunciation rules (called Tajweed); more than eight hours of speech were achieved [19-21]. Each audio file is a Quranic verse or a portion of it for long verses where the speaker must take a long breath.

In order to have a speech database useful for many goals, speech signals were manually and accurately segmented into three levels: word, phoneme and allophone. A new labeling system was proposed to annotate the speech segments [16] because the labeling systems available (e.g. IPA, SAMPA, BEEP, etc.) were not able to cover all Arabic sounds. However, the speech database consists of 44.1 KHz wav files of 16 millisecond utterances over its corresponding MFCC feature files, label files and TextGrids files.

Table I lists for each speaker, the number of sound files, their size and duration. The list of basic Arabic phonemes and their associated codes are shown in table II.

TABLE I. SOUND FILES AND THEIR DURATION BY SPEAKERS

Speaker Number	Speaker Initials	Number of Sound Files	Duration (minutes)	Size (MB)
1	AAH	600	49.36	249
2	AAS	590	52.09	261
3	AMS	612	45.78	229
4	ANS	597	49.72	250
5	BAN	585	54.75	276
6	FFA	578	44.11	220
7	HSS	601	49.76	251
8	MAS	580	46.24	232
9	MAZ	608	51.47	258
10	SKG	584	44.29	220
Total		5935	487.53 (8h, 8m)	2446

TABLE II. LIST OF BASIC ARABIC PHONEMES AND THEIR CODES

Arabic Orthography	Label	Arabic Orthography	Label
ـَ فتحة	as10	ص صاد	sb10
ـُ ضمة	us10	ض ضاد	db10
ـِ كسرة	is10	ط طاء	tb10
ء همزة	hz10	ظ ظاء	zb10
ب باء	bs10	ع عين	cs10
ت تاء	ts10	غ غين	gs10
ث ثاء	vs10	ف فاء	fs10
ج جيم	jb10	ق قاف	qs10
ح حاء	hb10	ك كاف	ks10
خ خاء	xs10	ل لام	ls10
د دال	ds10	م ميم	ms10
ذ ذال	vb10	ن نون	ns10
ر راء	rs10	هـ هاء	hs10
ز زاء	zs10	و واو	ws10
س سين	ss10	ي ياء	ys10
ش شين	js10	صامت	sil

In addition, the speech database contains a list of 60 Arabic phonemes, an Arabic dictionary, a list of all unrepeated words included in the whole eight hours speech database and other useful files needed for the recognizer development.

V. STATISTICS EXTRACTION METHOD

To extract statistics from the speech database, a computer program was designed using MATLAB programming language developed by MathWorks [22]. The occurrence probability of each basic phoneme, frequency of occurrence of basic phoneme, mean duration, Min and Max durations for each basic phoneme, mode and the median of duration for each basic phoneme were calculated. Durations are computed on the basis of phonemes boundary extracted from TextGrids files attached withal the speech database Sound.

These gathered statistics are displayed in Table 3 (see Table III) which also shows the labels used for every basic phoneme in the speech database. Fig. 1 shows the mean of basic phonemes durations measured in second. The frequency of each basic phoneme in the whole database is shown in Fig. 2. For an in-depth analysis of the collected statistic and for the purpose to have extra information about the characteristics of the basic Arabic phonemes, useful graphs are depicted in Figures 3, 4,5 and 6.

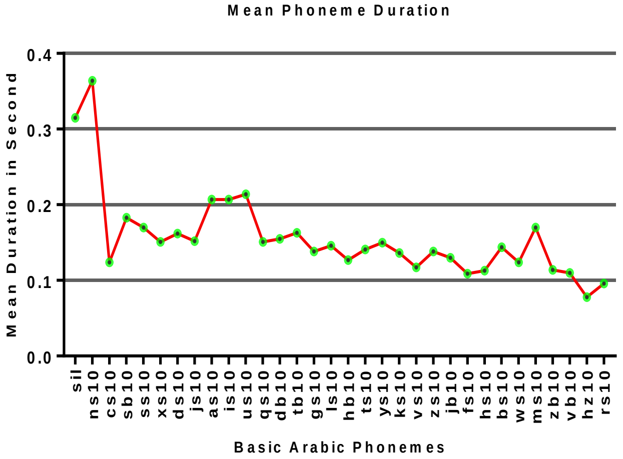


Fig. 1. Mean Duration of the Basic Arabic Phonemes

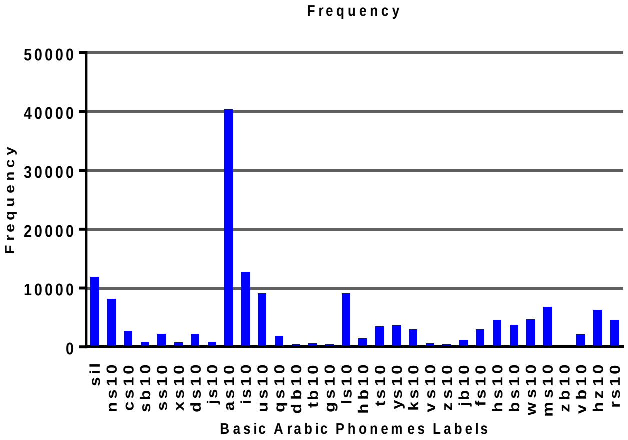


Fig. 2. Basic Arabic Phonemes Frequencies

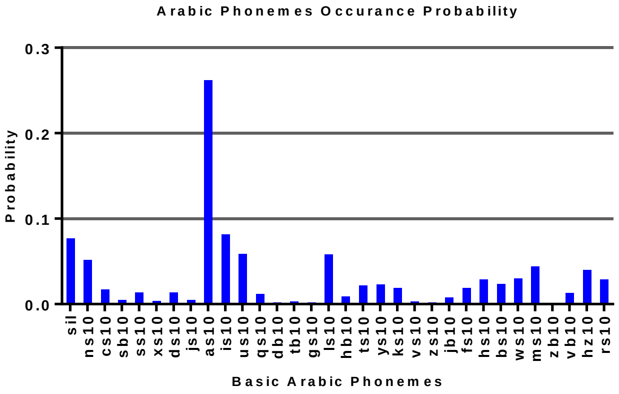


Fig. 3. Basic Arabic Phonemes Occurrence Probability

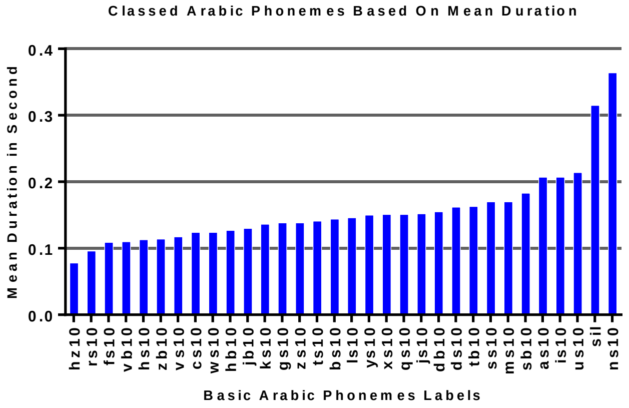


Fig. 4. Sorted Basic Arabic Phonemes based on their Means

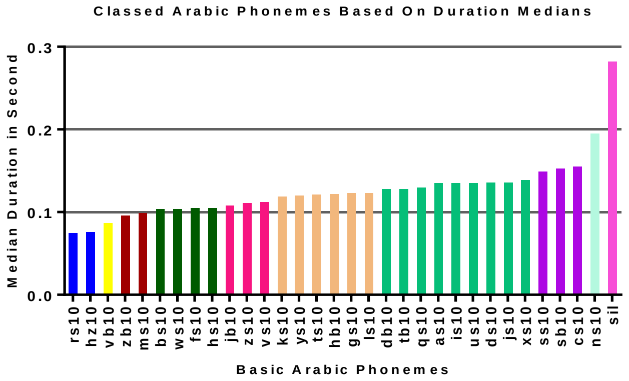


Fig. 5. Sorted Basic Arabic Phonemes based on their Medians

Fig. 3 shows the occurrence probability of the basic Arabic phonemes in the whole speech database. This useful graph will serve in defining the probability of missing phonemes during the decoding process. However, we noted that the phoneme “sil” denoting the silence regardless of its occurring places in the speech database is included in all depicted graphs.

In interesting outcome which is apparent from Fig. 4 proves that basic phonemes having equal or approximate mean values can be grouped into clusters. we assume that these clusters will be helpful for the purpose of enhancing performance of the baseline recognizer as we will evoke in the next sections. Basic phoneme duration medians give a clearly view of those clusters. Classes of the phonemes groups are being differentiated from each other and a clear parting among phoneme groups becomes more obvious, as seen in Fig. 5.

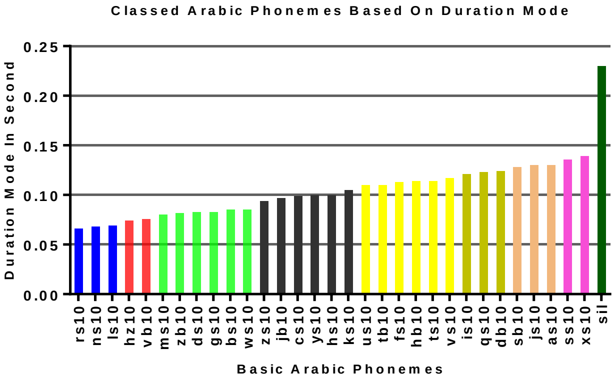


Fig. 6. Sorted Basic Arabic Phonemes based on their Modes

Another significant graph is the one demonstrating the most frequent duration value of all occurrences of a basic phoneme appearing in the "CA Sound Database". This is referred as the mode, and is displayed in Fig. 6.

VI. Statistics Analysis

When taking a look at the previous tables and graphs, we find that each basic phoneme occurs with various frequencies, the highest frequent ones are “as10” (فتحة), is10” (كسرة) and “us10” (ضمة), respectively, which designate the Arabic vowels. Otherwise the smallest frequent ones are “zb10” (حرف الظاء), “gs10” (حرف الغين), and “zs10” (حرف الزاء), respectively, ignoring the phoneme denoting the silence “sil” (صامت). From the results shown in Figures 2 and 3; it seems clear that when a phoneme is missed throughout the decoding process, phoneme "as10" is automatically the most probable one replacing it. Generally, the results concluded from Fig. 3 can be employed to correct the pronunciations for a misrecognized phoneme in spoken utterances during the recognition phase. The use of this information seems useful in enhancing the baseline system performance.

Fig. 4 illustrates the entire basic Arabic phonemes sorted on the basis of their average durations. From this Figure, we can clearly show the behavior of the basic phoneme durations through the whole speech database. Thus, the figure provides an explicit idea about the average duration of each phoneme, which means that a basic phoneme clusters being distinguished from it. For example, the basic phonemes “hz10” and “rs10” form the first cluster. The second cluster includes: “vb10”,”fs10” and “hs10”. The vowels form the last cluster in terms of the highest average durations. Usually, knowing the average length of a specific phoneme in a speech database can be utilized for estimating the appropriate number of the HMM states that represent it, which generally accelerate the estimation period and hence enhance the accuracy of recognition.

In Fig. 5 and Fig. 6, median and mode durations for each basic phoneme are displayed, where the basic phonemes clusters appear clearly. The outcomes of both figures could be helpful to make the correct decision in dealing with either misrecognized or missed phonemes. It means that replacing them with the near median or mode phoneme.

VII. Conclusion

In this paper, we have presented a collection of statistical data for Basic Arabic phonemes helpful in enhancing HMM-based automatic speech recognition systems performance. In the literature, the duration of phonemes is regarded as major distinctive feature characterizing the voice of a speaker. Knowing the duration of a particular phoneme in a spoken utterances can be utilized to estimate the length of the HMM chain describing it, which in consequence improves the system performance. These investigations were performed using a particular speech database of Quranic sounds including more than eight hours of speech and ten different male speakers. The numerical values are extracted using a computer program designed for this purpose. A discussion of these results with interpretations was also presented and reported graphically. Dividing phonemes into clusters on the basis of their median of the durations can help in decreasing the search for the appropriate phoneme during the decoding process, which in consequence increases system performance. Collected statistics provided can also be used to build or propose other techniques for phonemes classifications. While the probability distributions in HMM-based ASR systems are usually estimated with the Expectation-Maximization iterative algorithm, the statistics provided can be utilized as an initial condition for the estimation procedure, and, thus, speed up its execution time, or can also be utilized as a wanted model itself. We believe that the absence of necessary numerical data denoting, particularly, the basic Arabic phonemes behavior in classical Arabic language like those reported here gives an added value to the presented work. However, our future steps will focus on incorporating these statistics explicitly into HMMs in order to overcoming the classical HMM's weakness and, hence, improve HMM-based systems performance.

Acknowledgment

The presented work utilizes the results (Classical Arabic Sound Database) of a project previously funded by King Abed Al-Aziz City for Science and Technology (KACST) in Saudi Arabia under grant number “AT – 25 – 113”.

References

[1] D. Jurafsky and J. H. Martin, Speech and Language Processing, 2nd ed., Pearson Prentice Hall, 2009.

[2] G. Zweig and P. Nguyen, “A segmental CRF approach to large vocabulary continuous speech eecognition,” Proc.of IEEE ASRU, 2009.

[3] H. Sakoe, Two-level DP-matching - a dynamic programming-based pattern matching algorithm for connected word recognition, Readings in Speech Recognition, Morgan Kaufmann Publishers Inc, pp. 180-186, 1990.

[4] H. Jiang, “Discriminative training for automatic speech recognition: A survey,” Computer Speech & Language, Comput. Speech, vol. 24, no. 4, pp. 589–608, 2010.

[5] L. Deng and X. Li, "Machine Learning paradigms for speech recognition: An overview,” IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 5, pp. 1060–1089, May 2013.

[6] I, Oparin, Language Models for Automatic Speech Recognition Of inflectional Languages. PhD Thesis, University of West Bohemia, Plzen, Czech Republic (2009).

[7] Y.O.M. Elhadj, I.A. Alsughayeir, M. Alghamdi, M. Alkanhal, Y.M. Ohali, A.M. Alansari, Computerized teaching of the Holy Quran (in Arabic), Final Technical Report, King Abdulaziz City for Sciences and Technology (KACST), Riyadh, KSA,2012.

[8] Y.O.M. Elhadj, M. Alghamdi, and M. Alkanhal, “Phoneme-Based Recognizer to Assist Reading the Holy Quran,”, Recent Advances in Intelligent Informatics, Advances in Intelligent Systems and Computing, Springer, pp.141-152,2014.

[9] Y.O.M. Elhadj, M. Alghamdi, and M. Alkanhal, “Approach for Recognizing Allophonic Sounds of the Classical Arabic Based on Quran Recitations,”, Theory and Practice of Natural Computing, Lecture Notes in Computer Science, Springer, pp. 57-67, 2013.

[10] Y.O.M. Elhadj, Mohamed .O.M. Khelifa, A. Yousfi and M. Belkasmi. “An Accurate Recognizer for Basic Arabic Sounds,” ARPN Journal of Engineering and Applied Sciences, vol. 11, no. 5, pp. 3239- 3243, Mar. 2016.

[11] Mohamed O.M. Khelifa, Y.O.M. Elhadj, Y. Abdellah and M. Belkasmi, “Enhancing Arabic Phoneme Recognizer using Duration Modeling Techniques,”, in proc. of Fourth International Conference on Advances in Computing, Electronics and Communication - ACEC 2016, Dec 15, 2016, Rome-Italy.

[12] Mohamed O.M. Khelifa, Y.O.M. Elhadj, Y. Abdellah and M. Belkasmi, “An Accurate HSMM-based System for Arabic phonemes Recognition,” in proc. of The IEEE Ninth International conference on Advanced Computational Intelligence (ICACI 2017), Feb. 2, 2017, Doha, Qatar.

[13] S. Young, Large Vocabulary Continuous Speech Recognition: a Review, IEEE Signal Processing Magazine 13(5), pp. 45-57, 1996.

[14] Ali, A. et al., “A Complete KALDI Recipe for Building Arabic Speech Recognition Systems”, Spoken Language Technology Workshop (SLT), IEEE, 2014.

[15] Khalid, A. et al., "Arabic Phonemes Transcription using Data Driven,"The International Arab Journal of Information Technology, Vol. 12, No. 3, May 2015.

[16] Speaker-dependant continuous Arabic speech recognition. M.Sc. thesis, King Saud University, 2001.

[17] Hyassat H, Abu Zitar, “Arabic speech recognition using SPHINX engine,”, Int J Speech Tech 9(3–4):133–150, 2008.

[18] Azmi, M. et al., “Syllable-based automatic Arabic speech recognition in noisy-telephone channel,”, In: WSEAS transactions on signal processing proceedings, World Scientific and Engineering Academy and Society (WSEAS), vol 4, issue 4, pp 211–220, 2008.

[19] Y.O.M. Elhadj, M. et al., Design and Development of a High Quality Speech Corpus for Classical Arabic. Submitted for publication to the Language Resources and Evalauation Journal (LREV).

[20] Y.O.M. Elhadj, M. et al., Sound Corpus of a part of the noble Quran (in Arabic). Proc. of the International Conference on the Glorious Quran and Contemporary Technologies, King Fahd Complex for the Printing of the Holy Quran, Almadinah, Saudi Arabia, October 13-15, 2009.

[21] Y.O.M. Elhadj. Preparation of speech database with perfect reading of the last part of the Holly Quran (in Arabic). Proc. of the 3rd IEEE International Conference on Arabic Language Processing (CITAL'09), pp: 5-8, Rabat, Morocco, May 4-5, 2009.

[22] MATLAB and Statistics Toolbox Release 2013a The MathWorks, Inc., Natick, Massachusetts, United States.

TABLE III. THE BASIC ARABIC PHONEMES STATISTICS

Basic Arabic Phonemes	**Labels**	**Frequency of occurrence**	**Min duration in second**	**Max duration in second**	**Mean-duration in second**	**Mode**	**Median**	**Probability of occurrence**
صامت	sil	11875	0.022	8.576	0.315	0.230	0.282	0.077
نون	ns10	8160	0.021	1.458	0.364	0.068	0.195	0.052
عين	cs10	2700	0.033	0.420	0.124	0.099	0.155	0.017
صاد	sb10	838	0.079	0.388	0.183	0.128	0.153	0.005
سين	ss10	2175	0.071	0.384	0.170	0.136	0.149	0.014
خاء	xs10	770	0.072	0.420	0.151	0.139	0.139	0.004
دال	ds10	2190	0.039	0.433	0.162	0.083	0.136	0.014
شين	js10	867	0.080	0.478	0.152	0.130	0.136	0.005
فتحة	as10	40396	0.011	3.343	0.207	0.130	0.135	0.262
كسرة	is10	12755	0.030	1.833	0.207	0.121	0.135	0.082
ضمة	us10	9110	0.029	1.739	0.214	0.110	0.135	0.059
قاف	qs10	1870	0.080	0.792	0.151	0.123	0.130	0.012
ضاد	db10	443	0.021	0.629	0.155	0.124	0.128	0.002
طاء	tb10	560	0.073	0.464	0.163	0.110	0.128	0.003
غين	gs10	410	0.049	0.387	0.138	0.083	0.123	0.002
لام	ls10	9066	0.015	0.767	0.146	0.069	0.123	0.058
حاء	hb10	1457	0.050	0.335	0.127	0.114	0.122	0.009
تاء	ts10	3483	0.019	0.959	0.141	0.114	0.121	0.022
ياء	ys10	3677	0.019	1.392	0.150	0.100	0.120	0.023
كاف	ks10	3040	0.028	0.480	0.136	0.105	0.119	0.019
ثاء	vs10	600	0.032	0.311	0.117	0.117	0.112	0.003
زاء	zs10	440	0.060	0.352	0.138	0.094	0.111	0.002
جيم	jb10	1240	0.015	0.428	0.130	0.097	0.108	0.008
فاء	fs10	3020	0.016	0.369	0.109	0.113	0.105	0.019
هـاء	hs10	4559	0.029	0.376	0.113	0.100	0.105	0.029
باء	bs10	3739	0.012	0.654	0.144	0.085	0.104	0.024
واو	ws10	4647	0.016	1.021	0.124	0.085	0.104	0.030
ميم	ms10	6825	0.027	1.640	0.170	0.080	0.099	0.044
ظاء	zb10	176	0.054	0.360	0.114	0.082	0.096	0.001
ذال	vb10	2091	0.031	0.371	0.110	0.076	0.087	0.013
همزة	hz10	6281	0.008	0.295	0.078	0.074	0.076	0.040
راء	rs10	4620	0.014	0.403	0.096	0.066	0.075	0.029

Comparison of Discrete Cosine Transforms (DCT), Discrete Fourier Transforms (DFT), and Discrete Wavelet Transforms (DWT) in Digital Image Watermarking

Rosa A Asmara
Information Technology Department
State Polytechnics of Malang
Malang, Indonesia

Reza Agustina
Information Technology Department
State Polytechnics of Malang
Malang, Indonesia

Hidayatulloh
Information Technology Department
State Polytechnics of Malang
Malang, Indonesia

***Abstract*—Digital Image Watermarking is used recently to secure the image by embedding another digital image. It is typically used to identify ownership of the copyright of the signal. Frequency domain transformation methods used widely in Digital Image Compression and Digital Image Watermarking. They reduce the weakness of classics digital image watermarking such as Least Significant Bit (LSB) methods which is more noise-tolerant. Popular transformation method used are Two Dimensional Discrete Cosine Transform (2D DCT), Two Dimensional Discrete Fourier Transforms (2D DFT), and Two Dimensional Discrete Wavelet Transform (2D DWT). This paper will show the comparison result of those three transformation method. The experiments are comparison analysis of image watermark quality using Peak Signal to Noise Ratio (PSNR), color converting, image resizing, image optical scanning and the noise-tolerant of the image watermarked by giving Gaussian noise.**

Keywords—Digital Image Watermarking; 2D Discrete Cosine Transform (2D DCT); 2D Discrete Fourier Transform (2D DFT); 2D Discrete Wavelet Transform (2D DWT); Least Significant Bit method (LSB); Digital Signal Processing

I. Introduction

Digital Image Watermarking is used recently to secure the image by embedding another digital image. Komatsu and Tominaga are the first person using the term Digital Watermarking [1]. It is typically used to identify ownership of the copyright of the signal. The information are embedded in image is called a “digital image watermark”. The information where the watermark is to be embedded is called a “host image” [2,3]. Traditional method for Digital Image Watermarking used Least Significant Bit (LSB). Many researcher proposed the LSB method with some improvement and analysis to create better digital image watermark results [4,5,6]. The method for LSB will be explained in detail in Section 2.

LSB method for digital image watermarking has a weakness which cannot handle simple noise. Image watermarked also will loss the watermark information if some image processing is implemented such as Image Resizing and Image Cropping. Some Frequency Domain Transformations method is implemented to handle such weakness in traditional method. It is also proved that even the image is printed and scanned to return the format in digital, the watermark image can also be extracted smoothly.

Discrete Cosine Transform (DCT) are popular among science and engineering application, from image watermarking, steganography, and lossy compression for audio and image. Cosine function is used rather than sine function due to the critical for compression, fewer cosine functions are needed to approximate the typical signal. Cosine functions also express a particular choice of boundary condition in differential equations. DCT is similar to the Discrete Fourier Transform (DFT), but using only real numbers. DCT are equivalent of DFT of roughly twice the length, operating on real data with even symmetry and in some variants the input or output data are shifted by half a sample.

The remainder of this paper is organized as follows: Section 2 explained the 2D DCT, 2D DFT, 2D DWT and PSNR for analyzing the watermarked image. Section 3 describes our experiment and a technique for image watermarking in frequency domain information. Section 4 presents the experiments results. Conclusion and future work are given in the final section.

II. Proposed Method

A. Digital Image Watermarking

Digital image watermarking is one of the steganography branches, which is a technic to hide information in a digital media. The purpose is to protect important information [8]. The information inserted in digital image watermarking can be in text, image, or audio format file. There are some criteria to create good image watermarking:

- Imperceptibility: Good Watermark is invisible by human eye. One cannot distinguish well between original and watermarked image.
- Robustness: Good Watermark has to be resisted with file manipulation such as file compression, image noising, color converting, and image resizing.
- Security: Watermarked file can only be detected by the file owner or authorized ones.

- Recovery: Watermarked file must be converted back to original file. Main purpose of watermarking is for copyrighting file owner, which can be used for file authorization.

B. Least Significant Bit Method

Least significant bit (LSB) is very simple and less computation cost compare to the other method of transformations. Figure 1 shows the insertion method in LSB. For every bit starting from MSB of watermark image pixel is inserted to the LSB of some image source pixel. The results will be the addition of 1 if the watermark bit is 1 and same as previous value if the watermark bit is 0. Human eyes will be difficult to distinguishing the image source before and after watermarking. However, this method faces some image processing implementations such as Image Resizing and Image Cropping due to the pixel value modification. The PSNR of the LSB 1 bit substitution is 55.8784 [4].

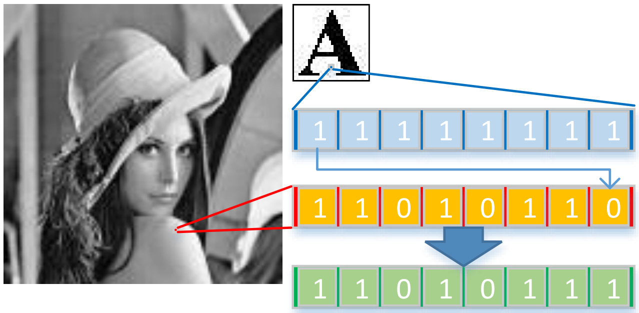


Fig. 1. LSB method of Digital Image Watermarking

C. 2D Discrete Cosine Transform (2D DCT)

DCT in image processing is first introduced by Ahmed, Natarajan and Rao [7]. DCT is similar to DFT, but using only real numbers. DCT turn over the image edge to make the image transformed into other form of even function. This is one of linear transformations in digital signal processing. 2D DCT is defined as:

$$F(jk) = a(j)a(k)\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} f(mn)cos\left[\frac{(2m+1)j\pi}{2N}\right]\cos\left[\frac{(2n+1)k\pi}{2N}\right] \quad [1]$$

The corresponding inverse discrete cosine transformation (2D-IDCT) is defined as:

$$f(mn) = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} a(j)a(k)F(jk)cos\left[\frac{(2m+1)j\pi}{2N}\right]cos\left[\frac{(2n+1)k\pi}{2N}\right] \quad [2]$$

The 2D-DCT can not only focusing on transform the main information of original image into the smallest low-frequency component, but also it can cause the image blocking effect being the smallest, which can realize the good compromise between the information centralizing and the computing complication, thus it obtains the wide spreading application.

D. 2D Discrete Fourier Transform (2D DFT)

Consider one N1 x N2 image, f(n1,n2), where we assume that the index range are n_1 = -M_1,…,M_1 and n_2 = -M_2,…,M_2, for mathematical simplicity, and hence N_1 = 2M_1 + 1 and N_2 = 2M_1 + 1. Let F(k_1,k_2) denote the 2D discrete Fourier Transform (2D DFT) of the image. F(k_1,k_2) are given by

$$F(k_1, k_2) = \sum_{n_1 n_2} f(n_1, n_2) W_{N_1}^{k_1 n_1} W_{N_2}^{k_2 n_2} = A_F(k_1, k_2) e^{j\theta_F(k_1,k_2)} \quad [3]$$

Where k_1 = -M_1,…,M_1, k_2 = -M_2,…,M_2, $W_{N_1} = e^{-j\frac{2\pi}{N_1}}$, $W_{N_2} = e^{-j\frac{2\pi}{N_2}}$, and the operator $\sum_{n_1 n_2}$ denotes $\sum_{n_1=-M_1}^{M_1}\sum_{n_2=-M_2}^{M_2}$, $A_F(k_1, k_2)$ is an amplitude component, and $e^{j\theta_F(k_1,k_2)}$ is a phase component.

The 2D Inverse Discrete Fourier (2D IDFT) of $F(k_1, k_2)$ is given by

$$f(k_1, k_2) = \frac{1}{N_1 N_2}\sum_{k_1 k_2} F(k_1, k_2) W_{N_1}^{-k_1 n_1} W_{N_2}^{-k_2 n_2}$$

Where $\sum_{k_1 k_2}$ denotes $\sum_{k_1=-M_1}^{M_1}\sum_{k_2=-M_2}^{M_2}$.

E. 2D Discrete Wavelet Transform (2D DWT)

Discrete wavelet transform (DWT) represents an image as a subset of wavelet functions using different locations and scales. It makes some decomposition images. Any decomposition of an image into wavelet involves a pair of waveforms: the high frequencies corresponding to the detailed parts of an image and the low frequencies corresponding to the smooth parts of an image. DWT for an image as a 2-D signal can be derived from a 1-D DWT. According to the characteristic of the DW decomposition, an image can be decomposed to four sub-band images through a 1-level 2-D DWT, as shown in Fig. 2. These four sub-band images in Fig. 4 can be mapped to four sub-band elements representing LL (Approximation), HL (Vertical), LH (Horizontal), and HH (Diagonal) respectively.

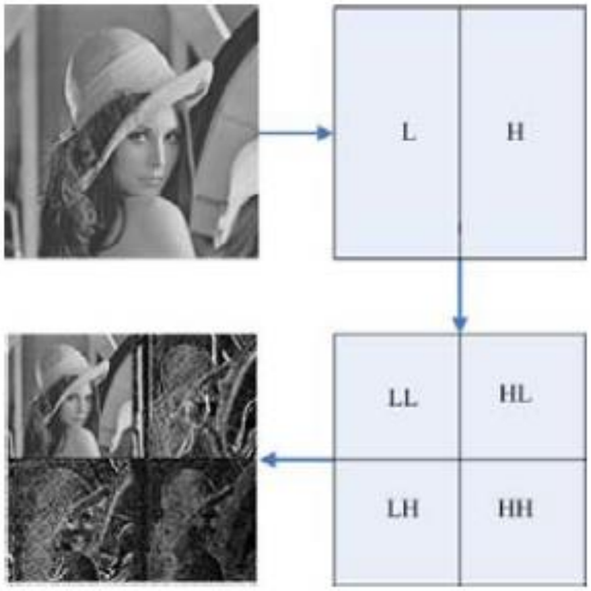


Fig. 2. 1-Level Decomposition 2D DWT

The discrete Wavelet Transform will decompose a given signal into other signal known as the approximation and detail coefficients. A given function f(t) can be expressed through the following representation:

$$f(t) = \sum_{j=1}^{L} \sum_{K=-\infty}^{\infty} d(j,K)\varphi(2^{-j}t - K) + \sum_{K=-\infty}^{\infty} a(L,K)\theta(2^{-L}t - K) \quad [4]$$

Where: φ(t) is the mother wavelet and θ(t) is the scaling function. *a*(L,K) is called the approximation coefficient at scale L and d(j,K) is called the detail coefficients at scale j. The approximation and detail coefficients can be expressed as:

$$a(L,K) = \frac{1}{\sqrt{2^L}} \int_{-\infty}^{\infty} f(t)\theta(2^{-L}t - K)dt \quad [5]$$

$$d(j,K) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} f(t)\varphi(2^{-j}t - K)dt \quad [6]$$

Based on the choice of the mother wavelet φ(t) and scaling function θ(t), different families of wavelets can be constructed.

F. Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR)

Peak signal-to-noise ratio (PSNR) is a ratio between maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR is usually expressed in terms of the logarithmic decibel scale. The signal is an original data, and the noise is the error from watermark system.

PSNR usually defined via the mean squared error (MSE). Given a noise-free m x n image I and noisy approximation K, MSE is defined as:

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (I(i,j) - K(i,j))^2 \quad [7]$$

The PSNR (in dB) is defined as:

$$\begin{aligned} PSNR &= 10log_{10}\left(\frac{MAX_I^2}{MSE}\right) \\ &= 20\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \\ &= 20\log_{10}(MAX_I) - 10\log_{10}(MSE) \quad [8] \end{aligned}$$

Where, MAX_I is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255.

III. Experiments

Figure below shows the general watermark image steps in frequency domain transforms.

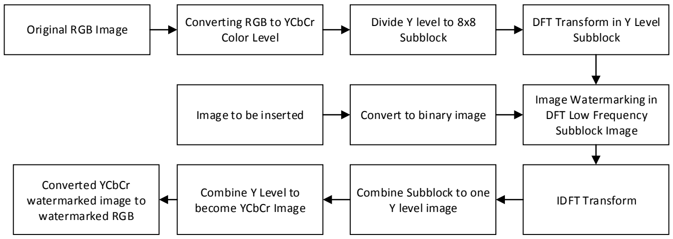

Original RGB Image
Converting RGB to YCbCr Color Level
Divide Y level to 8x8 Subblock
DFT Transform in Y Level Subblock
Image to be inserted
Convert to binary image
Image Watermarking in DFT Low Frequency Subblock Image
Converted YCbCr watermarked image to watermarked RGB
Combine Y Level to become YCbCr Image
Combine Subblock to one Y level image
IDFT Transform

Fig. 3. Watermark image steps in frequency domain transform

The usage of DCT, DFT, and DWT in image watermarking start by dividing image to 8x8 pixel sub-block. These sub-block will consist of 64 coefficient (1 DC-zero frequency coefficient and 63 AC coefficient in low frequency, middle frequency, and high frequency. Figure 4 shows the DCT frequency division.

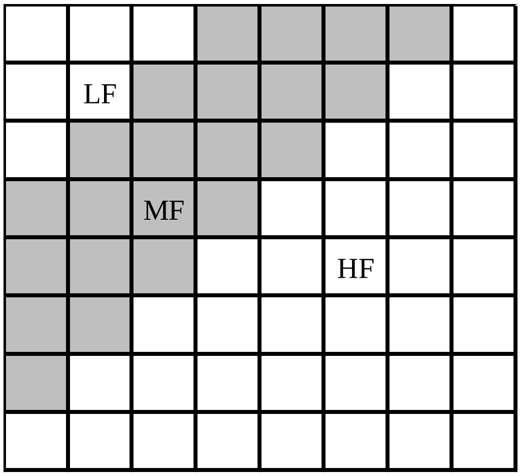


Fig. 4. DCT Frequency components

LF represents the low frequency, MF represents middle frequency, and HF represents sub-block highest frequency. In DCT transform, LF can be found in every edge and corner of image sub-block. In DCT and DFT, watermark bit image will be put in LF. In DWT, watermark bit image will be put in LH and HL. Figure 5 shows the diagram block for DCT watermarking process. This diagram block also represents for DFT and DWT.

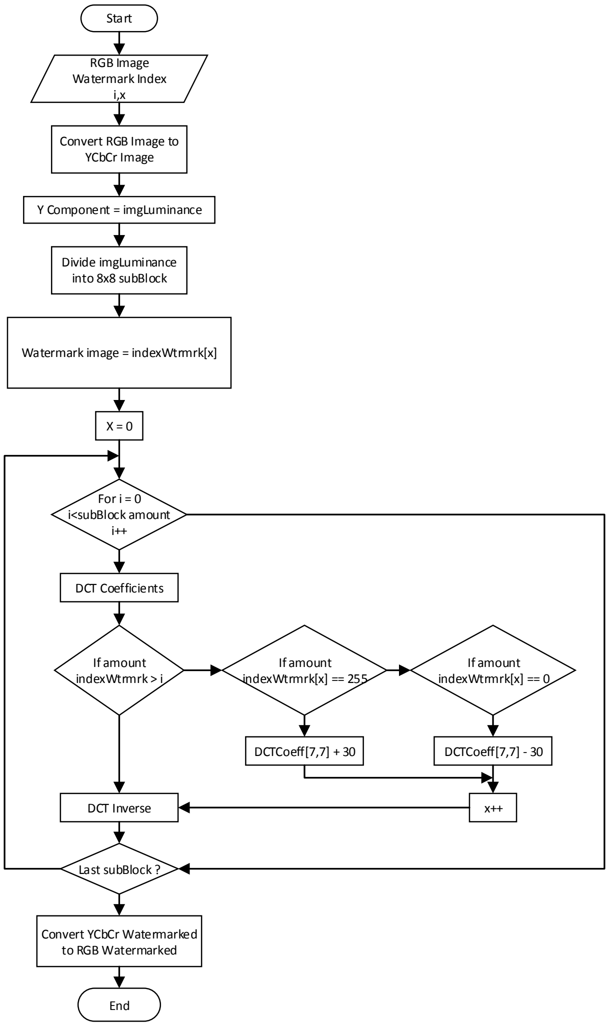

Start
RGB Image
Watermark Index
i,x
Convert RGB Image to YCbCr Image
Y Component = imgLuminance
Divide imgLuminance into 8x8 subBlock
Watermark image = indexWtmrk[x]
X = 0
For i = 0
i<subBlock amount
i++
DCT Coefficients
If amount indexWtmrk > i
If amount indexWtmrk[x] == 255
If amount indexWtmrk[x] == 0
DCTCoeff[7,7] + 30
DCTCoeff[7,7] - 30
DCT Inverse
x++
Last subBlock ?
Convert YCbCr Watermarked to RGB Watermarked
End

Fig. 5. Diagram blocks of DCT image watermarking

- *User* input *host image*, RGB image in which watermark image will be inserted. Watermark image will use binary image provided.
- RGB image will be converted to YCbCr color space, Y color space (luminance) component is the only component where the watermark image will be inserted, and thus the Y-component transformed using DCT, DFT, and DWT. Y-component is used because human perception more sensitive to the light intensity and the result for watermarked image will imperceptible. *Chrominance blue* (Cb) and *chrominance red* (Cr) component will not transform and will be used only for Inverse Transform.
- Y-component is divided into 8x8 sub-blocks.
- Each 8x8 sub-blocks then transforms using DCT, DFT, and DWT. Image watermark is a binary image with 0 and 1 pixel value. If the pixel value is 1 in the watermark image, then the sub-block host image in index [7, 7] will be add with 30, otherwise will be subtract with 30. This process is then repeated until the last watermark image pixel. The Y-component then detransformed using inverse-DCT, inverse-DFT, and inverse-DWT to get Y-component watermarked image.
- Y-component is combined again with Cb-component and Cr-component to get YCbCr watermarked image.
- Last proses is YCbCr watermarked converted again to RGB Space to get RGB watermarked.

IV. Results

We use 32x32 binary image resolutions as an image watermark. This image will embed on image with 65536 total pixel amount, since 1 pixel of binary image will embed on 8x8 image sub-block. Figure 5(a) is used as an image watermark and figure 5(b) as an original image.

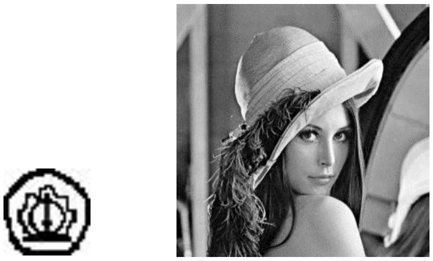

(a) (b)

Fig. 6. (a) Image Watermark, (b) Original Image

Table 1 shows the results of embedding image watermark to the LF, MF, and HF of original image.

TABLE I. PSNR Results of Image Watermark in DCT, DFT, and DWT

Freq.: Embed Pixel Coordinate	DCT PSNR (dB)	DFT PSNR (dB)	DWT PSNR (dB)
LF: [w0, h1]	40,24	32,2	33,78
MF: [w3, h3]	39,96	35,18	33,87 (HL) 33,87 (LH)
HF: [w7, h7]	40,82	32,18	33,87

TABLE II. PSNR Results of Gaussian Noise Attack in Watermark DCT, DFT, And DWT

20 % Gaussian Noise to the watermarked image			
Frequency	PSNR DCT	PSNR DFT	PSNR DWT
LF	19,36	19,04	18,90
MF	19,26	19,00	19,01 (HL) 18,95 (LH)
HF	19,29	19,11	19,15

TABLE III. PSNR Results of Compression Attack in Watermark DCT, DFT, and DWT

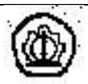
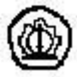
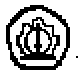
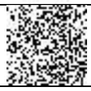
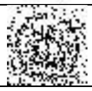
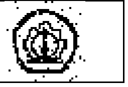
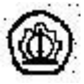
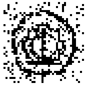
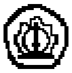
Compression test using RIOT Application			
Level	Extraction DCT	Extraction DFT	Extraction DWT
25 %			
50 %			

TABLE IV. PSNR RESULTS OF IMAGE CONTRAST MANIPULATION IN WATERMARK DCT, DFT, AND DWT

HF Watermarked on Image Contrast Manipulation			
Image after Contrast Manipulation	PSNR DCT	PSNR DFT	PSNR DWT
 50% Contrast Addition	27,57 	21,48 	20,32 

V. CONCLUSION

Frequency domain transformation methods are used widely in Digital Image Compression and Digital Image Watermarking. It reduces the weakness of classics digital image watermarking such as Least Significant Bit (LSB) methods which is more noise-tolerant. Popular transformation method used are Two Dimensional Discrete Cosine Transform (2D DCT), Two Dimensional Discrete Fourier Transforms (2D DFT), and Two Dimensional Discrete Wavelet Transform (2D DWT). This paper proposed the comparison between those three transformation methods. The experiments are image watermark quality analysis using Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR), and the noise-tolerant of the image watermarked by giving Gaussian noise in it. The experiments shows result of:

- DCT Transformation embed on High Frequency is the best for image watermarking. It has 40.82 dB PSNR values, as shown in table 1.
- Adding 20% Gaussian noise to the watermarked image, the best Transformation for Gaussian Attack is DCT watermark in Low Frequency. It has 19.36 dB PSNR value, as show in table 2.
- Compressing image using RIOT application for all three transformation, DWT shows the best results as shown in table 3.

Using image contrast manipulation, DCT shows the best results as shown in table 4.

REFERENCES

[1] Bender, W., Gruhl, D., Morimoto, N. and Lu, A(1996).: Techniques for data hiding. IBM Systems Journal, vol. 35, nos. 3&4.

[2] Saraju Prasad Mohanty(,January 1999)“Watermarking of Digital Images”, Submitted at Indian Institute of Science Bangalore, pp. 1.3 – 1.6,.

[3] Katzenbeisser, S. and Petitcolas, F(1999).: Information hiding techniques for steganography and digital watermarking. Artech House Books.

[4] Puneet Kr Sharma and Rajni(2012).: ANALYSIS OF IMAGE WATERMARKING USING LEAST SIGNIFICANT BIT ALGORITHM, International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.4, July 2012.

[5] Abdullah Bamatraf, Rosziati Ibrahim and Mohd. Najib Mohd. Salleh(2011).: A New Digital Watermarking Algorithm Using Combination of Least Significant Bit (LSB) and Inverse Bit, Journal of Computing Vol.3, Issue 4, April 2011, ISSN 2151-9617.

[6] Deepshikha Chopra, Preeti Gupta, Gaur Sanjay B.C., Anil Gupta (2012).: Lsb Based Digital Image Watermarking For Gray Scale Image, IOSR Journal of Computer Engineering (IOSRJCE) Vol.6, Issue 1, pp 36-41, Sep-Oct 2011, ISSN 2278-0661.

[7] Ahmed, Natarajan, and Rao (1974) : On Image Processing and a Discrete Cosine Transform, IEEE Trans. On Computer C-23(1): 90-93

AUTHOR PROFILE

Rosa A. Asmara: Lecturer of Information Technology Department in State Polytechnics of Malang, Indonesia. Research interest are Image Processing and Computer Vision, Pattern Recognition, and Artificial Intelligence.

Reza Agustina, Hidayatulloh: Undergraduate student in Information Technology Department, State Polytechnics of Malang, Indonesia.

Web Application Development by Applying the MVC and Table Data Gateway in the Annual Program Budget Management System

A. Medina-Santiago, A. Cisneros-Gómez, E. M. Melgar-Paniagua
Center of Investigation, Development and Innovation Technology
University of Science and Technology Descartes/Institute Polytechnic National
Tuxtla Gutierrez, Chiapas, Mexico

G. B. Nango-Sólis, E. A. Moreno-López, M. E. Castellanos-Morales, D. B. Cantoral-Díaz, L. M. Blanco-Gonzalez
Dept. of computer and Dept. Finances
Institute Technology of Tuxtla Gutierrez
Tuxtla Gutierrez, Chiapas, Mexico

***Abstract*—This paper is the result of the development of the Web application to register the Annual Work Program, in which goals and actions are assigned the financial resources to manage the annual work program identified. In this paper, we have identified five types of users: the first is the Administrator, in charge of monitoring the goals programmed in the period, as well as the resource assigned to reach those goals; the second corresponds to the purchasing department who is in charge of contacting the supplier and at the same time inform financial and warehouse of the acquisition through the system; the third corresponds to a warehouse in charge of validating the material and generate entry / exit official vouchers and send the purchase order to financiers; the fourth user corresponds to financial, this will identify through the system that all the procedure is completed to make the payment; and finally, the fifth user make up the set of all remaining departments. Finally, the system presents flexibility in case it is necessary to go adding departments.**

***Keywords*—WEB Applications; MVC; Data Gateway Table; Software engineering; incremental; iterative**

I. Introduction

Now-a-days, the systems for administering the Annual Operational Program (AOP) are very useful systems that facilitate the programming, and administration of the financial resources assigned to each department, as well as establishing the following guidelines to reach what is scheduled in the year for an Institution. The AWP allows monthly activities to be monitored to meet the goals in the year [1], while the AOP is responsible for budgeting the financial resources necessary to fulfill the AWP [2, 3].

Information systems allow you to manage and control a company. Whenever an efficient recording of the data is carried out, all the information necessary to prepare reports, records and, consequently, make a decision will be available and handy. Information systems are composed of computer equipment, human resources, data that are entered into the system, programs that are executed on the computer to produce results, telecommunications and procedures that include policies and operating rules, which interact between Yes to achieve the organization's objectives [4].

At the Technological Institute of Tuxtla, Gutierrez, the system was developed with the purpose of speeding up the purchase process, because there are projects and needs that must be addressed promptly [5], but due to the administrative process that must follow, implies that pass through several departments before completing the above procedure. The departments that are involved are: Planning, Material Resources and Services (in the Purchasing and Warehouse offices), financial Resources, as well as the department requesting the acquisition of the good or service. In each of these departments, formats are generated that are captured in Word or Excel as the case may be. In most of the cases, the data to be captured are redundant data that, although systematize, the retrieval of the document should be automatic.

Currently, in the Technological Institute of Tuxtla, Gutierrez, year by year he carries out the annual work program (AWP), through actions programmed by strategic and key processes. So, the purpose of this project is a web application that allows the registration of the AWP, the resources allocated each of the actions of the strategic process through the Annual Operational Program (AOP), the registration of the expenses to carry out such actions. The registration of purchases, the entry of material, product or service through warehouse and the notice to financial resources to release the payment, as well as a series of reports in each of the elements already mentioned.

To develop the system, it was considered to each of the users that intervene to follow up a purchase process, the user (department) requesting a service or product must first have the resource to be able to prepare the corresponding requisition, it should be clarified that the user at all times manages to see the tracking of the acquisition from the moment he request until he arrives at the warehouse.

This work has the purpose of optimizing the time in the release of the resource for the acquisition of the product or service, as well as the systematization of the entire process, so that in a future work can be made efficient decision making based on historical data.

To achieve this, an application with the Software engineering methodology has developed, in this case allowed the development planning. As part of the analysis phase is used case of use of the UML (Unified Modeling Language) diagrams, to then define modeling of business processes via BPMN diagram to identify the workflow, as you can see more later.

II. Methods

For the development of the system, the agile incremental Iterative methodology was used [6, 7], see figure 1, where each increment or new feature added to the project consists of planning. Design, construction, testing and commissioning.

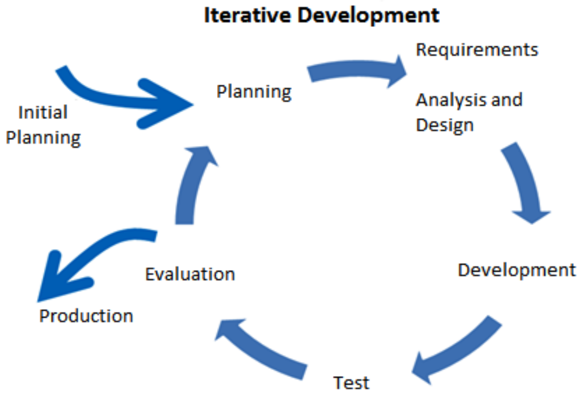


Fig. 1. Iterative Methodology

The reason why this methodology was applied is because it presents a great flexibility to make changes and / or to add new characteristics, with witch it is possible that in each iteration a version of the fully functional system is generated (see Figure 2).

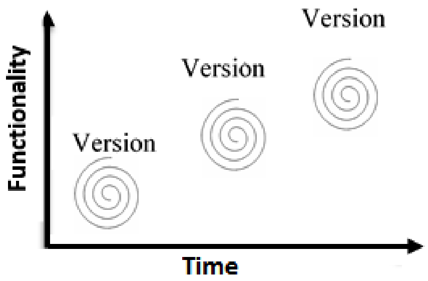


Fig. 2. Graph of increase of functionality

A. Collection of Data

As part of the first iteration, we reviewed how we work the Annual Work Plan (AWP), how departments are involved when defining their annual work plan; Base on the information obtained, the web interface that integrates with the ITTG Integral Information System (SII) was developed. In this interface, there is an Administrator who adds the goals to be met, can generate reports and view the information captured from each Department, there are also Departments represented by the department role they can select the goals to which they contribute.

In the second iteration, the Annual Operational Plan (AOP) was known which consist of making the department budget to consolidate each goal selected by each department. Based on the way operations for the AOP are performed in the current administration, a web interface has been developed that integrates with the SII which has two users identified. Administrator and Department, the administrator is in charge of establishing the limit Budget, generate reports an approve applications among other tasks, while the Department can in this section make the budget tor he current period, in addition to making the following request: Request for Requisition, Request for services an Request for Viaticum; Each request after being captured can be formatted in a PDF document for later printing.

B. Used tools

PHP is a programming language generally used in the programming of dynamic, open source and platform independent websites, fast, with a large library of functions and a lot of documentation [8].

For the development of the application Zend Framework 1.12 is used which is supported by the version of PHP that is used in the production server [9]; This framework uses the MVC (Model-View-Controller) methodology for the organization of the source code, in addition it uses the design pattern Table Data Gateway that allows a perfect separation of SQL code of the rest of the classes of the classes of the system [10, 11].

C. MVC (Model-View-Controller)

The MVC considers there operating instances: the controller and the view, in figure 3 the operating scheme is presented [12].

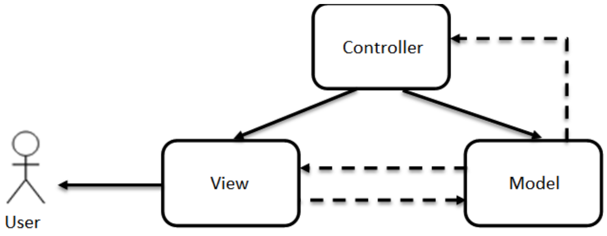


Fig. 3. Representation of the operation of the MVC

- **Model:** An object that represents some information about the domain. This is a non-visual object that contains all data and behaviors different from those used by the user interface. In its most pure OO (Object Oriented) form the model is an object within the Domain Model [13]
- **View:** Defines exactly what is presented to the user. Normally the controllers pass the data to each view to be represented in the view with some format. Views also collect user data, with forms.
- **Controller:** The Controller integrates the pattern, always. Manipulate the model, decide what I will show based on the user's request and other factors.

D. Table Data Gateway

It is a pattern that usually provides sufficient separation, without the cost of using a proxy pattern; this is also known as Data Access Object (DAO), this pattern uses a specialized mask for each type of object that we want to store in the database [11]. The corresponding abstract classes are presented in figure 4.

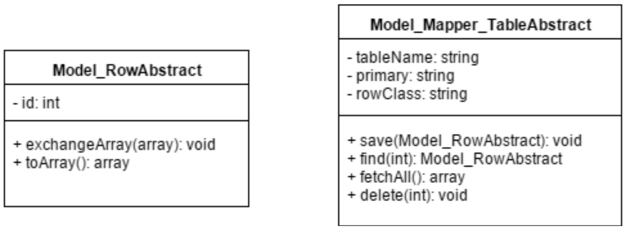


Fig. 4. Graph of increase of functionality

III. Development

It began with the planning of activities in each of the phases of the Software Development Lifecycle, to the give way to the stage of requisition analysis and design. Here, the system's activities were identified from the point of view of each user, making it possible to clearly establish two users that are fundamental for the good operation of the system, on the one hand the administrator and on the other the clients of the system, in this case as the Heads of Department of ITTG. The figure 5 below shows the activities diagram for carrying out the AWP and the allocation of the corresponding financial resource.

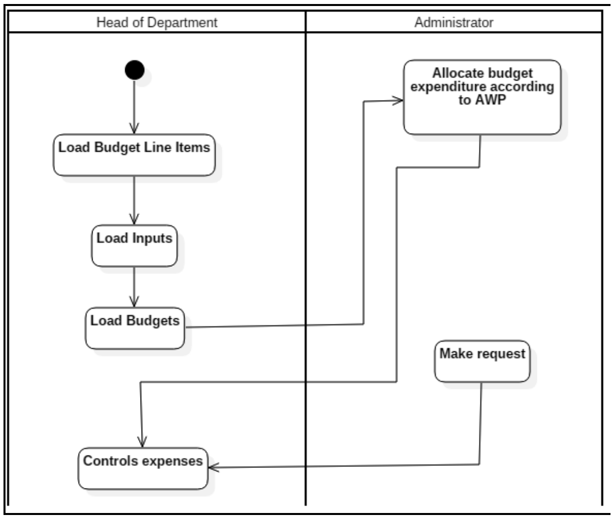


Fig. 5. Activity Diagram

In order to make a purchase and/or payment to the supplier or service provider, the following activities were identified:

- Department, make request.
- Administrator (Dept. planning, programming and budgeting), makes the assessment of the application and can authorize.
- Purchases (within the Department of Materials Resources) authorize, rejects and can print formats.
- Warehouse receives products if these have been requested from a supplier who seals the invoice received to pass the apartment. Of Material Resources and authorized the purchase.
- Financial Resources, check where the requisition is and wait for it to be in physical format in the department.

On the other hand, the administration for de Department was identified and developed. Resource Materials consisting of two parts:

- The first one is Purchases, where the authorization of the Request is made, if there is a purchase of materials, the request is made to suppliers, if the is no need to turn to Financial Resources directly the payment order.
- The second is Warehouse, is responsible for receiving the materials requested from the supplier and seals the invoice received for delivery to purchase and can generate the purchase order for Warehouse to generate: entry /exit vouchers and offices payment.

A monitor was integrated to know in which department the office is located and the status (accepted, rejected or canceled) of the document, this is presented to all users with their respective permissions so a department will display the status of their accepted applications and Financial Resources will identify in which part of the process each application is located.

Finally, the Financial Resources department receives the payment order and invoice, in order to issue the payment and thus finalize the process.

As part of obtaining system requirements and analysis, it was also possible to identify the functionality of the system from the point of view of the user [14-16], as can be seen in figure 6 to 9.

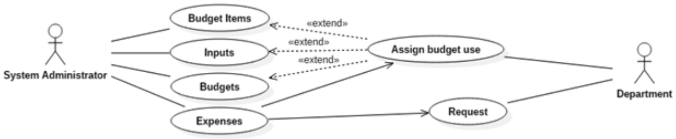


Fig. 6. Functionality of the System Represented with use case diagrams

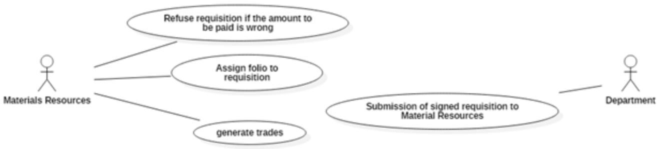

Refuse requisition if the amount to be paid is wrong
Materials Resources
Assign folio to requisition
generate trades
Submission of signed requisition to Material Resources
Department

Fig. 7. Interaction of the material resources department with the system

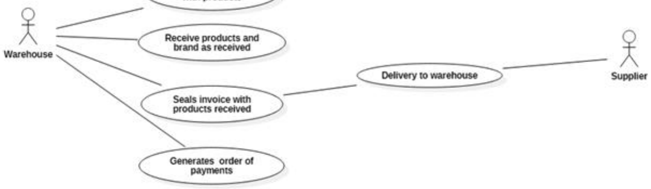

New acquisitions with products
Warehouse
Receive products and brand as received
Seals invoice with products received
Generates order of payments
Delivery to warehouse
Supplier

Fig. 8. Warehouse Interaction with Supplier

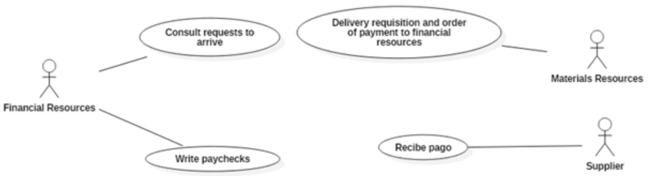

Consult requests to arrive
Delivery requisition and order of payment to financial resources
Materials Resources
Financial Resources
Write paychecks
Recibe pago
Supplier

Fig. 9. Interaction between Financial Resources, Material Resources and the Supplier

As part of the process of identifying the process that are performed in the departments, when a request to the AOP was made, the following steps were identified in the diagram of figure 10 BPMN (Business Process Model and Notation) [17].

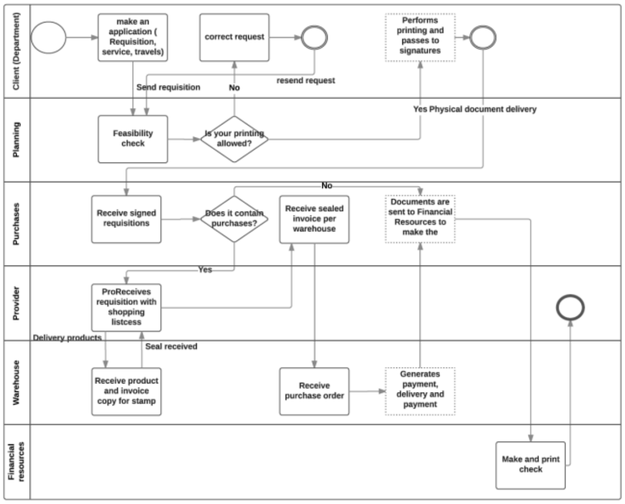

Client (Department)
make an application (Requisition, service, travels)
correct request
Performs printing and passes to signatures
Send requisition
No
resend request
Planning
Yes Physical document delivery
Feasibility check
Is your printing allowed?
Purchases
No
Receive signed requisitions
Does it contain purchases?
Receive sealed invoice per warehouse
Documents are sent to Financial Resources to make the
Yes
Provider
ProReceives requisition with shopping listcess
Delivery products
Seal received
Warehouse
Receive product and invoice copy for stamp
Receive purchase order
Generates payment, delivery and payment
Financial resources
Make and print check

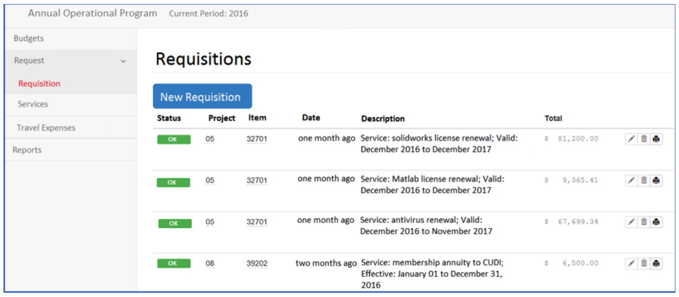
Fig. 10. BPMN Diagram (Business Process Model and Notation)

In the design of the system the classes that would later intervene in the development of the system were identified, see appendix A.

For the storage of the information was created the Database that can be seen in the Appendix B.

IV. Results

The main product that is obtained from this is obtained from this Project involves the following users: Client (Departments), Department of Planning, Purchasing Office, Office of Warehouse and financial resources. It is the customer who makes the request of a requisition (see Figure 11)



Requisitions

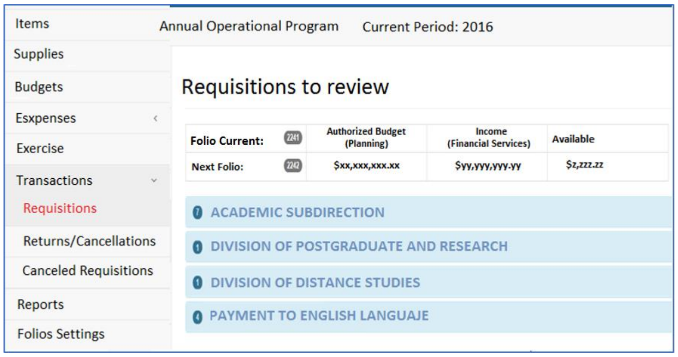
Status	Project	Item	Date	Description	Total
OK	05	32701	one month ago	Service: solidworks license renewal; Valid: December 2016 to December 2017	$ 81,200.00
OK	05	32701	one month ago	Service: Matlab license renewal; Valid: December 2016 to December 2017	$ 9,365.41
OK	05	32701	one month ago	Service: antivirus renewal; Valid: December 2016 to November 2017	$ 67,699.34
OK	08	39202	two months ago	Service: membership annuity to CUDI; Effective: January 01 to December 31, 2016	$ 6,500.00

Fig. 11. Capture, edit, delete, and printing of requisition (purchase order)

The Department of Planning is responsible for performing the checking of feasibility and authorize the requisition, relying on the interface that is presented in Figure 12, the tables allows you to view the following information:

- Folio Current: is the amount of requisitions that have already been approved so far.
- Authorized budget (Planning): represents in money the authorizations of requisitions.
- Income (Financial Services): Displays the amount of money up to which you can exercise in the year, it should be noted that the latter data can be modified several times by the Department of financial resources, due to perceive income through the year.
- Available: reports the amount of money that you can still get to spend in the current fiscal year and is the result of real income (financial) - Authorized Budget (Planning).

Immediately after the table shows a list of departments with a number on the left side means the amount of requisitions that a department has requested and are still pending authorize.



Requisitions to review

Folio Current: 2241	Authorized Budget (Planning)	Income (Financial Services)	Available
Next Folio: 2242	$xx,xxx,xxx.xx	$yy,yyy,yyy.yy	$z,zzz.zz

7 ACADEMIC SUBDIRECTION

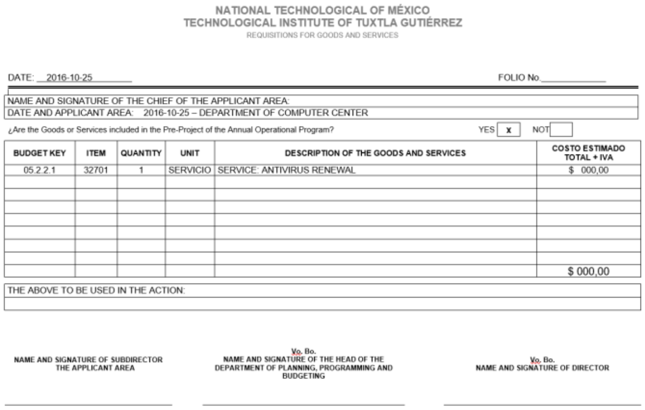
1 DIVISION OF POSTGRADUATE AND RESEARCH

1 DIVISION OF DISTANCE STUDIES

1 PAYMENT TO ENGLISH LANGUAJE

Fig. 12. Requisition information pending authorize, as well as the amount of money pending exercise

Once approved the requisition can be printed on the system, the customer will see the legend authorized to the select the print icon as shown in figure 11, to print the requisition as seen in Figure 13.



NATIONAL TECHNOLOGICAL OF MÉXICO
TECHNOLOGICAL INSTITUTE OF TUXTLA GUTIÉRREZ
REQUISITIONS FOR GOODS AND SERVICES

DATE: 2016-10-25 FOLIO No.

NAME AND SIGNATURE OF THE CHIEF OF THE APPLICANT AREA:
DATE AND APPLICANT AREA: 2016-10-25 – DEPARTMENT OF COMPUTER CENTER

¿Are the Goods or Services included in the Pre-Project of the Annual Operational Program? YES ☒ NOT ☐

BUDGET KEY	ITEM	QUANTITY	UNIT	DESCRIPTION OF THE GOODS AND SERVICES	COSTO ESTIMADO TOTAL + IVA
05.2.2.1	32701	1	SERVICIO	SERVICE: ANTIVIRUS RENEWAL	$ 000,00
					$ 000,00

THE ABOVE TO BE USED IN THE ACTION:

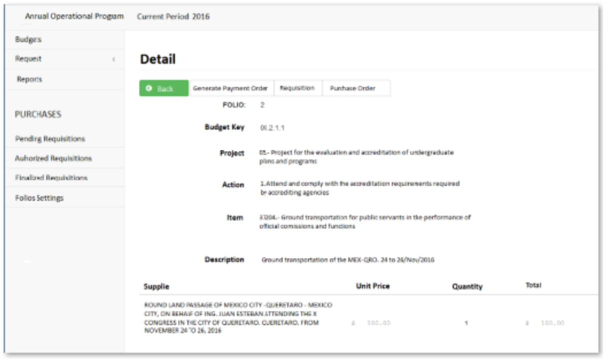
NAME AND SIGNATURE OF SUBDIRECTOR THE APPLICANT AREA

Vo. Bo.
NAME AND SIGNATURE OF THE HEAD OF THE DEPARTMENT OF PLANNING, PROGRAMMING AND BUDGETING

Vo. Bo.
NAME AND SIGNATURE OF DIRECTOR

Fig. 13. Requisition

On the other hand, the purchasing office, taking the requisition authorized access to view the detail of the requisition (see figure 14), where you can generate purchase order to send to the supplier, as well as the payment order to send it to the Department of financial resources (see Figure 15).



Detail

FOLIO: 2

Budget Key 06.2.1.1

Project 05.- Project for the evaluation and accreditation of undergraduate plans and programs

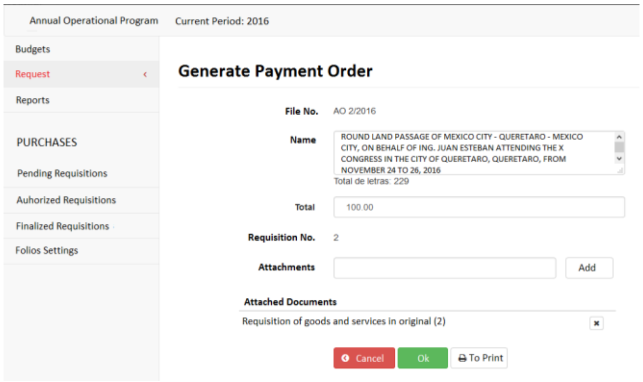
Action 1.Attend and comply with the accreditation requirements required by accrediting agencies

Item 3704.- Ground transportation for public servants in the performance of official comissions and functions

Description Ground transportation of the MEX-QRO. 24 to 26/Nov/2016

Supplie	Unit Price	Quantity	Total
ROUND LAND PASSAGE OF MEXICO CITY -QUERETARO - MEXICO CITY, ON BEHALF OF ING. JUAN ESTEBAN ATTENDING THE X CONGRESS IN THE CITY OF QUERETARO, QUERETARO, FROM NOVEMBER 24 TO 26, 2016	$ 100.00	1	$ 100.00

Fig. 14. Detail of requisition in the account of the purchasing office



Generate Payment Order

File No. AO 2/2016

Name ROUND LAND PASSAGE OF MEXICO CITY - QUERETARO - MEXICO CITY, ON BEHALF OF ING. JUAN ESTEBAN ATTENDING THE X CONGRESS IN THE CITY OF QUERETARO, QUERETARO, FROM NOVEMBER 24 TO 26, 2016

Total de letras: 229

Total 100.00

Requisition No. 2

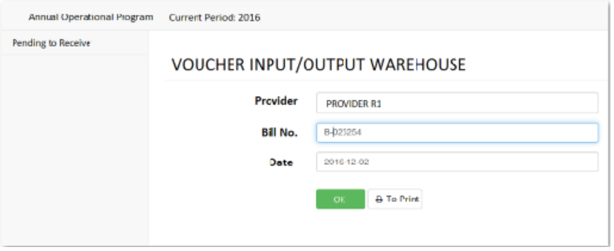
Attachments

Attached Documents

Requisition of goods and services in original (2)

Fig. 15. Summary of the Purchase Order

In the Store account is responsible for generating the vouchers input/output (see figure 16 and 17). Which does the person who delivers the well as well as for the one who receives it, sign.



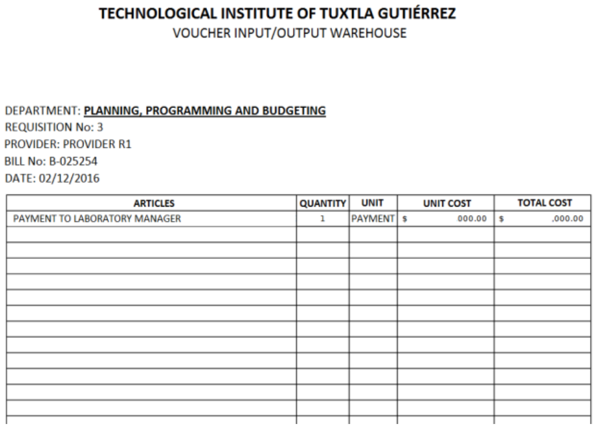
VOUCHER INPUT/OUTPUT WAREHOUSE

Provider PROVIDER R1

Bill No. B-025254

Date 2016-12-02

Fig. 16. Option to generate vouchers input/output



TECHNOLOGICAL INSTITUTE OF TUXTLA GUTIÉRREZ
VOUCHER INPUT/OUTPUT WAREHOUSE

DEPARTMENT: PLANNING, PROGRAMMING AND BUDGETING
REQUISITION No: 3
PROVIDER: PROVIDER R1
BILL No: B-025254
DATE: 02/12/2016

ARTICLES	QUANTITY	UNIT	UNIT COST	TOTAL COST
PAYMENT TO LABORATORY MANAGER	1	PAYMENT	$ 000.00	$,000.00

Fig. 17. Voucher of entry and exit of warehouse

The systematization of the administration of the presupposed annual, has allowed have a better control in the expenditure, is has reduced in a 50% the time in that takes in authorize is a buy since this is requested, also allows get of way immediate indicators of the State of has of each Department, the budget concentrate by split budget and program institutional, breakdown of the budget of investment with charges to income own between others.

V. Discussion

The importance of the development of this system allows the carrying out of proper control of the projects established by the Annual Operational Program, determine the goals and actions and consistently identify the allocation of financial resource, generate the requisition orders, authorization of the same and go after deducting the resources allocated by budget headings.

With specific records of dates, preparation authorization and monitoring within the purchasing process.

Concerning the results obtained by the development of the web application, the procedure of purchases that account the institute has been of great contribution, since that will allow for greater control and follow-up of the same, particularly stresses the importance of budgetary control that by law should be not to make expenditures outside the authorized budget; Take to the track in the whole process of shopping cart until the payment of the same enables all users to have knowledge of the times and to be able to identify if according to their responsibility and competence has been made.

It is considered as a comprehensive system that benefits thorough the adequate fulfilled, comparison of information in regard to the budget allocated and the authorized, generates information for the analysis of budgetary control and compare the financial information with the areas involved in the process of shopping, Department of Planning, authorizes and validates the requisition, Department of material resources that performs the quotations, runs the purchase, generates payment order and finally conclude the process with the payment of the purchase generated in the department of Financial Resources.

On the other hand, from the point of view of software engineering using BPMN diagrams in development, allows to express the process of business for the systematization of processes, MVC allows you to divide the logic of the design business doing more scalable to the project, while Table Data Gateway encapsulates the database access in a natural way, using an object that acts as gateway to a table of database, of this way a change of version in the database can help to the code source of the version previous follow running.

VI. Conclusions

The present work is very useful, allows to optimize the time in the process of the application until the acquisition and payment of the good or service. The system is intuitive and achieves the easy operation for each of the users, obtaining quickly a series of reports that are required in all administrative procedures, as well as, easily store and retrieve the registration of the procedures.

The Web application follows the process and notifies through flags in which status the acquisition process is. The application has the flexibility to add new department and users at the beginning of the period.

Maintenance of this system in the future will be an easy task, since bases that will help to achieve it have been established from the beginning, through the division of the business logic to implement the model-view - controller together with the encapsulation of data by means of Table Data.

As part of a future work, is referred to include advanced electronic signature as part of automation and digitalization of processes.

On the other hand, must analyze historical data for optimization of expenditure, through "data warehouse" to improve the quality of decision-making. It earlier will allow identify what are the services or inputs that are bought and in what time of the year, it is acquired, as well as who are the departments that it requested. The data warehouse also will identify cross purchases that are made.

Also, it aims to apply Dataminig with the purpose of identifying trends or predict the increase of the expenditure for the following exercise annual.

Acknowledgment

The authors thank and credit the National Technological of México / Technological Institute of Tuxtla Gutiérrez for all the facilities and support for the realization of this project in the Departments of Computer Center, Planning, Programming and Budgeting, Material and Financial Resources.

References

[1] ITTG. Procedure of the ITTG Quality Management Systems, 2015.

[2] Juan, J, Gutierrez, J., Garcia, I., Ramirez, A., Baró, J., Pozas, J., López, A., and Vilvhis, A. “Conservation and management of a protected natural area of the valley of México”, Colegio de Ciencias Geográficas del Estado de México, México, 2013.

[3] TecNM. Annual Operational Program (AOP). Available online: http://www.tecnm.mx/programacion-presupuestal/program-operativo-annual-poa. 2015.

[4] Cohen, D., Asin, E., “Information Systems For Business”, McGraw-Hill, Mexico, D.F. 2000, pp.3-26.

[5] Chuayffet-Chemor, E., Quintero-Quintero. M., Mendez-Navarro, J. L., “Institutional Program of Innivation and Development 2013-2018 of the Technical Institute of Tuxtla Gutierrez”, National Technology of Mexico, Mexico, 2013, pp.64-66.

[6] Sommerville, I., “Software engineering”, Pearson-Addison Wesley, Spain, 2005, pp. 358-378.

[7] Constantine, L.L. and Lockwood, L.A.D., “Software for use: A Practical Guide to the Models and Methods of Usage-Centered Design”, Pearson Education, USA, 1999.

[8] Perz, C., “MySQL for Windows and Linux”, Alfaomega Ra-Ma, Mexico, 2008.

[9] Allen, R., Lo, N., Brown, S., “ Zend Framework in Action”, Manning (MAEP – Manning Early Access Program), 2009, pp. 9.

[10] Zend framework: User Guide – Database and models.

[11] ADOdb – Database Abstraction Layer for PHP.

[12] Pitt, C., “Pro PHP MVC”, Appress, 2012.

[13] Krasner, G.E., Pope, S.T. , “A description of the Model-View-Controller User Interface Paradigm in the Smalltalk-80 System. J. Object Oriented Progra”, 1988.

[14] Rumbaugh, J., Jacobson, I., Booch, G., “The unified modeling language reference manual”, Addison-Wesley, USA, 2001.

[15] Fowler, M., Scott, K., “UML distilled”, Pearson Education, Mexico, 1999.

[16] Rosenberg, D., Scott, K., “Use case driven object modeling with UML: A practical approach”, Addison-Wesley, USA, 2001.

[17] White, S. A. and Miers, D., “BPMN modeling and reference guide”, Future strategies, 2008.

Semantic Sentiment Analysis of Arabic Texts

Sana Alowaidi, Mustafa Saleh, Osama Abulnaja
Computer Science Department
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Twitter considered as a rich resource to collect people's opinions in different domains and attracted researchers to develop an automatic *Sentiment Analysis* (SA) model for tweets. In this work, a semantic *Arabic Twitter Sentiment Analysis (ATSA)* model is developed based on *supervised machine learning (ML)* approaches and semantic analysis. Most of the existing Arabic SA approaches represent tweets based on the *bag-of-words (BoW)* model. The main limitation of this model is that it is semantically weak; where words considered as independent features and ignore the semantic associations between them. As a result, synonymous words that appear in two tweets are represented as different independent features. To overcome this limitation, this work proposes enriching the tweets representation with concepts utilizing *Arabic WordNet (AWN)* as an external knowledge base. In addition, different concepts representation approaches are developed and evaluated with *naïve Bayes (NB)* and *support vector machine (SVM)* ML classifiers on an Arabic Twitter dataset. The experimental results indicate that using concepts features improves the performance of the ATSA model compared with the basic BoW representation. The improvement reached 4.48% with the SVM classifier and 5.78% with the NB classifier.**

Keywords*—*Arabic Sentiment Analysis; Twitter; Semantic Relations; Arabic WordNet; Machine Learning

I. Introduction

Currently, Twitter is considered to be one of the most popular microblogs. It has allowed people to communicate, share comments, and express their opinions on almost all aspects of daily life at an increasing rate. Since analyzing huge volumes of opinionated text remains a formidable task, the high demand for automated sentiment analysis (SA) models became a necessity.

Sentiment analysis, which is also called opinion mining, is the computational study of people's opinions, sentiments, and attitudes about topics, entities, people, and events, that are expressed in texts [1]. It aims to assign a predefined sentiment class to online texts as negative, positive, or neutral. SA plays a substantial role in several domains such as financing, marketing, politics, and social.

One of the main approaches used to solve the SA problem is the supervised machine learning (ML) approach. In this approach, texts are represented by feature vectors which are used to train ML classifiers, such as naïve Bayes (NB) and support vector machines (SVMs), to infer a combination of particular features yielding a certain sentiment class. The resulting classifier model is then used to predict the sentiment class of the new un-annotated documents [1].

The performance of the SA model relied on the classifier algorithm and the text representation model. Various classifiers have been adopted for SA, but the challenging task is to engineer a set of powerful features to build a good representation model [1]. The vector space model (VSM) [2], also called the bag-of-words (BoW) model, is considered as a fundamental text representation model used in most ML approaches because of its simplicity and effectiveness. This model represents texts as a weighted features vector with words as basic features.

Many of the existing approaches in both the English and Arabic languages attempt to enhance the performance of the SA model by expanding the BoW model with different features such as word n-grams, POS tags, and stems. Also, new microblog features were proposed for Twitter data. However, the resulting representation models still suffer from a common limitation, they are semantically weak. The BoW model considers the words as independent features and ignores the semantic associations between the words. For example, it treats synonymous words as unrelated features. Moreover, in this model, only words that are explicitly mentioned in the training dataset are used to train the classifier, thereby ignoring the words in the testing documents that were not found in the training documents.

Recently, several studies, most of which were in the English language, have been proposed using a new semantic concepts representation model in various text mining (TM) fields including clustering [3], topic classification [4-6], and SA [7, 8]. Rather than representing the documents in their lexical space depending on BoW features, the semantic approach represents the documents in their semantic space as a set of concepts features extracted utilizing an external knowledge base (KB) such as WordNet (WN).

The Arabic language is a Semitic language which consists of 28 letters. It is a cursive language, in which word formation consists of connecting letters to each other. As opposed to the English language, Arabic writing starts from right to left, and has no capitalization.

The Arabic language is one of the fastest-growing languages on the web with about 168 million Arabic-speaking people using the Internet [9]. According to the Internet World Stat 2016 ranking [9], the Arabic language ranked in the top five languages used most on the Internet. However, while much SA research has been done for the English language, since it is a dominant language of science, little has been done for the Arabic language. The Arabic language poses a number of challenges, especially in regards to sentiment analysis. It not only that it has a very complex morphology compared to the

English language, but it is also a very derivational and inflectional language which makes morphological analysis a very complicated task [10, 11].

This paper presents an Arabic Twitter Sentiment Analysis (ATSA) model, a semantic sentiment analysis model for Arabic Twitter data using ML approaches. Unlike existing Arabic SA models which represent tweets texts in their lexical space based on BoW features, semantic concepts representation approach was proposed which aims to represent tweets in their semantic space by taking into account the semantic relationships between the words by utilizing the Arabic WordNet (AWN).

The rest of this paper is structured as follows. Section II examines previous related work. Section III describes the ATSA model. Section IV presents the proposed concepts representation approach in detail. Section V discusses experimental settings and results. The last section concludes the paper and gives directions for future work.

II. Related Works

Recently, different Arabic SA models based on ML approaches have been proposed with various features and classifier algorithms for social media and microblogging services. Most of the proposed approaches represent documents based on the BoW model, and try to extend word features with different features such as n-grams (e.g., bi-gram, tri-gram), stems and POS tags.

Shoukry and Rafea proposed a sentiment classification for Arabic tweets [12]. They investigated using different sets of n-gram features with SVM and NB classifiers. Duwairi and Qarqaz in [13] built a SA model for Arabic Twitter and Facebook comments. In their model, the texts were represented as a set of word bi-gram features. They also investigated the effect of using term frequency (TF) and term frequency-inverse document frequency (TF-IDF) weighting schemes with SVM, NB, and K-nearest neighbours (K-NN) classifiers. Abdul-Mageed et al. in [14] presented a subjectivity and sentiment analysis system (SAMAR) based on a SVM classifier for different Arabic social media applications: Web forums, chat, Wikipedia Talk Pages, and Twitter. They studied different features including word n-grams, POS tagging, and word stems. Also, many stylistic features related to social media applications were investigated. The results showed that the classifier performance relied on the type of the dataset and features used.

Duwairi [15] proposed a SA approach for Arabic tweets written in Jordanian Arabic dialectical and Modern Standard Arabic (MSA). The researcher suggested improving the performance of the SVM and NB classifiers by transforming words in tweets from their dialect form to MSA. Hammad and Al-awadi in [16] focused on studying SA on Arabic hotel reviews collected from Twitter, Facebook, and YouTube. They employed NB, VSM, DT and back-propagation neural network (BPNN) ML classifiers with BoW, POS tag and stem features. The results showed that among the classifiers, the SVM classifier achieved the best average accuracy, followed by NB, DT and finally BPNN. Elghazaly et al. [17] evaluated the use of two classifiers, SVM and NB, on the SA of Egyptian political election tweets. The tweets were represented using BoW features with a TF-IDF weighting scheme. The results showed that the NB classifier achieved better accuracy and a faster time than the SVM.

Despite the efforts made in previous approaches, they still suffered from a common limitation: they were semantically weak. They ignored the semantic relationship between words in the documents. Different areas of text mining in the English language, such as text clustering, topic classification and SA, have recently seen an increase in research in an attempt to cope with the BoW model's limitation. To cope with the limitation, they built semantic text representation models that incorporate semantic concepts as features using an external KB, such as WN or named entity tools.

Hotho et al. [3] are considered among the first to propose a semantic representation using WN concepts as features for clustering fields. Three representation strategies were suggested: 1. Add concepts (AddC) as extra features to the BoW model. 2. Replace words with their concepts (ReplC). 3. Use bag-of-concepts (BoC) features only. Different word sense disambiguation (WSD) strategies were followed: selecting the first concept (FstC), all concepts (AllC), and disambiguation by context. The TF-IDF weight was applied with a k-means clustering algorithm. The experiments showed that using the semantic WN concepts features were promising and outperformed the baseline BoW model. Also, Baghel and Dhir in [18] proposed a hierarchy clustering algorithm to cluster the documents based on the concepts representation. The concepts were extracted from WN using the FstC WSD strategy. The TF-IDF weighting scheme was used. The proposed approach achieved better performance than traditional approaches.

A concept-based representation approach for topics-based classification of news articles was proposed by Elberrichi et al. [4]. The proposed approach utilized WN concepts to represent documents via various representation strategies: add concepts as extra features to the BoW model, replace concepts with words, and use BoC features only. Two WSD methods were used, FstC and AllC. The classifier model applied the TF-IDF weight with cosine distance similarity. The experiments showed that using the semantic WN concepts features with AddC and FstC WSD methods outperformed the baseline BoW model.

In the SA field, Balamurali et al. [8] proposed using WN concepts features to represent texts in travel reviews datasets. Two incorporation strategies were used, AddC and BoC, with two WSD methods, manual and automatic. They used the SVM classifier and found that using the AddC strategy with the manual WSD method achieved the best performance with an accuracy of 90.20%, which increased the performance of the SA by 5.3 % over the baseline BoW.

Gautam and Yadav in [7] proposed a semantic WN synonyms analysis method for SA on a Twitter dataset. The approach relied on checking the semantic synonym similarity between words in the testing and training tweets datasets. If a synonym similarity between the words was found, words in the testing data would be replaced with their synonyms in the training data. The approach was evaluated using different ML classifiers. The results showed that the NB classifier with TF weight obtained the superior results compared to the other classifiers used, SVM and maximum entropy (ME).

Another significant approach for SA on Twitter data was proposed by Saif et al. [19]. The approach utilized the semantic concepts, extracted from named entity tagger tools, as an additional feature into a training dataset for SA. The approach was based on the idea that specific entities and concepts tend to have a more consistent correlation with positive or negative sentiments. Knowing these correlations helps determine the sentiment of semantically relevant entities, even if those entities never appeared in the training set. They used the TF weight schema and NB classifier. The proposed approach outperformed the baseline feature BoW with POS.

III. Arabic Twitter Sentiment Analysis Model

A SA model for Arabic Twitter data based on the ML approach was developed. The overall architecture of the ATSA model consisted of two main phases, training and testing. In the training phase, the classifier needed to learn from a set of labeled tweets. It was then used to classify unlabeled tweets in the testing phase. Each phase consisted of the following steps: text preprocessing, features extraction, and classification. The general process of the ATSA model is illustrated in Fig. 1. First, the tweets datasets needed to be collected and annotated. After that, the tweets were preprocessed to eliminate the noise. Then, the features representation model was constructed. This step is critical because the type of extracted features and the manner in which they are built influences the performance of the ML classifier. Two different types of features were extracted, BoW and semantic concepts features, which were used to build the texts representation models. Finally, the ML classifier is trained and evaluated on unlabeled data. This section discusses the ATSA steps in more detail.

A. *Text Preprocessing*

Text preprocessing is an essential step in microblog data to clean the input tweets and eliminate noise and unnecessary data [20, 21]. Preprocessing consisted of the following steps: adding tags, data cleaning, normalization, tokenization, and stop words removal.

1) Adding Tags

In Twitter, people express their sentiments using different emoticon symbols. Thus, the emoticon symbols in tweets require special handling. Furthermore, some of the punctuation marks, such as exclamation mark (“!”) and question mark (“?”), are related to people’s emotions [22]. In this step, emoticons symbols are replaced with their corresponding meaningful word tags that represent their sentiment. Examples of the used emoticons are displayed in Table 1.

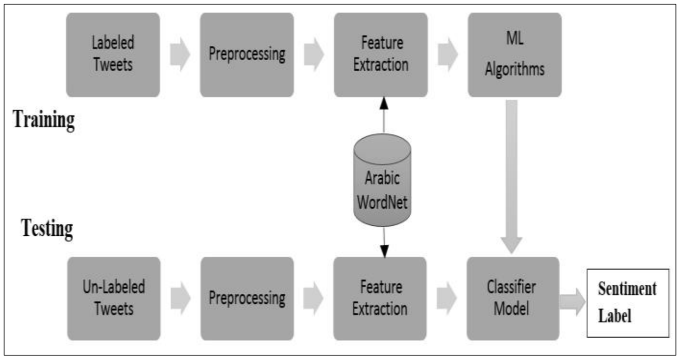


Fig. 1. Main ATSA Phases

2) Data Cleaning

Data cleaning is a critical task for dealing with the noisy nature of Twitter data. This step consisted of removing items from tweets that do not include any sentiments. As such, the following items were removed: URLs, re-tweet (RT) entities, usernames, numbers, single Arabic letters, non-letter characters (e.g., + = % $), and punctuation marks except question marks and exclamation marks (e.g., . ,: “” ; ').

3) Normalization

The normalization task is important in order to produce consistent word forms. Normalization for the Arabic text consisted of the following steps:

- Stripping diacritics: e.g. " الْعَرَبِيّة " to “ العربية ”.
- Stripping lengthening (Tatweel): e.g. “العربيـــــة ” to “العربية ”
- Removing "ال" from the beginning of words: e.g."العربية" will be "عربية"
- Replacing the letter "ة" with "ه"
- Replacing the letter "ى" with "ي"
- Replacing the letters "أ-إ-آ" with "ا"
- Normalizing repeated letters: e.g. “سعاااادة ” to “سعادة”

4) Tokenization

In this step, the tweet text was split into a sequence of tokens where each token represents a single word based on whitespaces.

5) Stop Words Removal

Removing stop words is a common step in text preprocessing. Stop words (such as from, in and of) are very common words that are frequently repeated in the dataset and do not provide any useful information to the text analysis. Removing these words allows the focus to be on the more important words and helps in dimension reduction. A list[1] of stop words was extended with many Arabic informal dialect words such as “كذا, عشان". The stop words from the list, except for negations, were removed.

TABLE I. EXAMPLES OF THE EMOTICONS WITH THEIR TAGS

Symbols	The Tags
“:)”, “:-)”, “:)”	HAPPY
“:(”, “:-(”, “: (”	SAD
!	Exclamation
?	Question

B. Features Extraction

Supervised ML algorithms require an appropriate representation of the documents as a features vector. The vast majority of ML approaches use the VSM [2], where each document is represented as a weighted features vector.

Different text representation models were created for tweets based on two extracted features: BoW and semantic concepts. The features need to be weighted using the term frequency-inverse document frequency (TF-IDF) [20] weighting scheme. This scheme helps reduce the weight of the features that appear in multiples dataset documents. It is defined as:

$$\text{TF-IDF}(f_n, di)=\text{TF}(f_n, di)\ .\ \text{IDF}(f_n) \tag{1}$$

where $\text{TF}(f_n,d_i)$ is the frequency of the feature f_n, $\text{IDF}(f_n)$ is defined as:

$$\text{IDF}(f_n)=\log\frac{|D|}{DF(f_n)} \tag{2}$$

where $\text{DF}(f_n)$ refers to the number of documents in D that include the feature f_n. The |D| is the total number of documents in the dataset.

1) The Bag-of-Words Representation

The BoW model used in most text mining applications has been shown to be quite effective in the SA field. To build the feature vectors, it considers the words as basic informative aspects of the texts. It consists of distinct words that appear in the dataset after preprocessing the tweets. Rather than depending on word features only, emoticon symbols are used as extra features with the BoW model to indicate the sentiment of the Arabic texts.

2) Concepts Representation

Representing tweets with BoW models neglects the semantic associations between words. As a result, synonymous words that appear in two tweets are represented as different independent features, and the model would not detect any related features between the tweets. This work proposed representing the tweets in their semantic space by incorporating semantic concepts to the tweets’ features space. This helps classify the sentiment of tweets that did not mention any words found in the training dataset, but did contain similar synonymous words. The concepts representation approach is described in detail in Section IV.

C. Classification

In this step, the resulted representation is supplied to the ML classification algorithm to build and learn a classifier model from training labeled tweets that can predict the sentiment label of new unlabeled tweets. Various supervised ML classifiers have been applied in previous research work on SA. The ATSA model was evaluated using the most common algorithms NB and SVM.

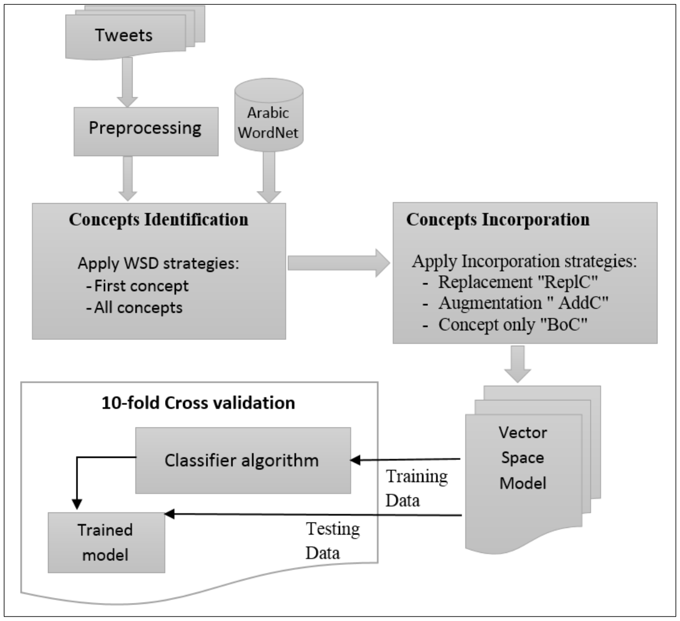


Fig. 2. Semantic ATSA architecture

IV. CONCEPTS REPRESENTATION APPROACH

Arabic WordNet (AWN) [23] is a lexical and semantic recourse of the Arabic language based on the English Princeton WordNet. It semantically groups words together into concepts based on their meaning. A concept, also named a synset, is a basic object in the WordNet to express a set of synonym words that share at least one sense.

The proposed concepts representation approach depends on extracting and employing semantic concepts features utilizing AWN, as shown in fig 2. To develop the concepts representation model, two main steps are required:

- Identify the concepts features.
- Incorporate the concepts using different strategies.

A. Concepts Identification

The target of this task is to identify the concepts in tweets by utilizing the AWN ontology. To extract the concepts, a number of steps are performed. First, words in each tweet are mapped to their concepts. The WordNet returns an ordered list of all related concepts. They are ordered from most appropriate to least appropriate. Then, for words that have many concepts, it is important to select the most appropriate meaning using the WSD strategy. While building an advanced WSD approach is beyond the scope of this research, the research concentrated on simply determining whether the WSD strategy was needed to produce a good performance. The simplest WSD strategies that

[1] Available From: https://code.google.com/p/stop-words/

were used in previous works [3, 4, 7, 8, and 18] were applied; they are first concept (FstC) and all concepts (AllC) strategies.

- First Concept (FstC): This strategy selects the first concept from the returned list as the disambiguation method.
- All Concepts (AllC): It is considered a basic strategy that selects all concepts of the word from the returned concepts list.

B. Concepts Incorporation

In this step, the extracted semantic concepts from tweets are incorporated as extra features to represent the tweets. It was proposed to use different incorporation strategies (augmentation "AddC", replacement "ReplC", and concept only "BoC") which have been previously developed in [3, 4, and 8] for English text mining applications.

1) Augmentation "AddC"

This strategy augments the identified concepts in the tweets into the BoW model as additional features with their corresponding words. By using the "AddC" strategy, the tweets are represented by all the extracted concepts and all the tweet's words. In this strategy, the size of the features is enlarged by the semantic concepts, and the new size is defined as $|F'|= |F| + |C|$. Where $|F'|$ is the total number of features, $|F|$ is the primary feature size, and $|C|$ is the number of semantic concepts associated with the words.

2) Replacement "ReplC"

This strategy replaces all words with their mapped concepts identified in the tweets. By using the "ReplC" strategy, tweets represented by concepts and words which have no map concepts in AWN. This strategy helps in reducing the features space, where the new size is defined as $|F'|= |F| - |W_c| + |C|$, where $|W_c|$ is the total number of individual words that are substituted by concepts.

3) Concept Only "BoC"

In this strategy, tweets are represented by their extracted concept only without any of their words. By using the BoC strategy, the size of the feature space is the same as the extracted semantic concepts $|F'|= |C|$.

V. Experiments and Evaluation

In this work, different tweet representation approaches were experimented to determine the best approach that improved the performance of ATSA model. The proposed semantic concepts representation against the basic BoW features was compared. Moreover, the variations of the semantic representations that resulted from applying the BoC only, AddC, and ReplC strategies with AllC and FstC WSD methods were evaluated. All of the experiments were conducted with two ML classifiers, SVM and NB, using RapidMiner[2] , which is a popular data mining tool.

A. Arabic Twitter Dataset

An Arabic Twitter corpus was built for SA by collecting tweets regarding people’s sentiments in different domains (politics, sports, social and companies). To automate the process, a python script was implemented to collect data from the Twitter API[3] using the Tweepy[4] library. The collection process was based on certain hashtags that represent important events or topics for each domain.

After collecting the tweets, each tweet was manually annotated with a sentiment class label as positive or negative. In this step, two human annotators were asked to read the tweets and assign sentiment labels to them. Most of the time, they agreed about the sentiment label. When they disagreed, a another human annotator was asked to determine the final label. The final dataset consisted of about 826 tweet text documents consisting of 413 positive and 413 negative tweets.

B. Evaluation Method and Performance Measurements

The performance of the developed approach was evaluated using F-measure. It is the harmonic mean of precision and recall. Precision and recall are two standard evaluation metrics widely used to evaluate the effectiveness of classification algorithms on a given category [24, 25].

The Precision (P), is the number of correctly classified positive tweets divided by the number of tweets labeled as positive by the system. It is defines as:

$$\boldsymbol{P} = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3}$$

The Recall (R), is the number of correctly classified positive tweets divided by the number of positive tweets in the dataset. It defines as:

$$\boldsymbol{R} = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{4}$$

Given P and R, the F-measure is defined as:

$$\mathbf{F} = 2\,.\frac{P *\ R}{P + R} \tag{5}$$

To ensure reliable results, all of the experiments were conducted using a ten-fold cross validation method [24, 25].

TABLE II. F-Measure Value of NB and SVM Classifier with All Representation Approaches

features	**NB**		**SVM**	
BoW	85.99		91.15	
	AllC	***FstC***	***AllC***	***FstC***
Concepts only	85.96	89.14	87.04	93.7
Concepts (AddC)	89.72	91.77	92.59	95.63
Concepts (ReplC)	86.79	88.93	90.87	95.11

[2] https://rapidminer.com
[3] https://dev.twitter.com
[4] http://www.tweepy.org

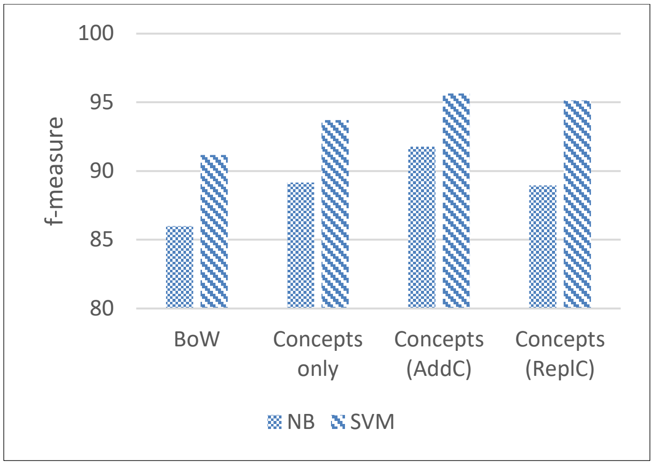


Fig. 3. The values of the F-measure for NB and SVM classifier with different representation approach and FstC WSD

sResults

Table 2 displays the results of different tweets representation with the NB and SVM ML classifiers. The semantic representation model using AWN concepts features was found to help improve the accuracy of the ATSA model. That is because the representation of the text was enriched with the semantic concepts feature which helped preserve the semantic relations between the words, such as synonyms, and produce more common concepts features that identify the related sentiment class. . For example, “فرح” (happy) and “مسرور” (glad) are synonymous words and both carry positive sentiments, the synonym relation between the words can be preserved only if the words were treated as concepts, not just as independent words.

As shown in fig 3, the performance of all of the classifiers improved with all of the proposed concepts representations. Furthermore, using the SVM classifier was found to outperform the NB classifier in almost all representation models. The highest F-measure value reached 95.63% when using the AddC concepts representation with the FstC WSD methods and SVM classifier.

Also, from the results, it is clear that using the AddC incorporation strategy provides the best performance over all concepts representation approaches. The BoC only representation discards the words that do not appear in the AWN. So, in this case, it may lose some of the distinctive word features which represent the sentiment class.

Moreover, regarding the effect of WSD, it is obvious from the experimental results that using simple FstC WSD outperformed the AllC method with almost all concepts representation, as illustrated in fig 4. Thus, using all the concepts could produce some noise data and mislead the sentiment classification.

VI. Conclusion and Future Works

In this work, the effect of using semantic AWN concepts features to represent tweets on the proposed ATSA model was demonstrated. Various approaches were proposed for building the concepts representation model using BoC only or combining the BoW with the concepts following two strategies: concept augmentation and concept replacement.

The experiments showed that using concepts features outperforms the baseline BoW model and opens great opportunities to build a robust SA model for Arabic tweets. Furthermore, among all of the approaches, augmentation concepts representations with the FstC methods achieved the best accuracy.

For future, the researchers plan to examine the semantic concepts representation model on larger datasets. Also, the conducted experiments proved that using a simple WSD method had a good effect on the concepts representation. Thus, developing more advanced WSD methods is critical for the Arabic language. Moreover, the researchers suggest developing an approach for extracting the concepts features from Wikipedia and using them to extend the representation of Twitter data.

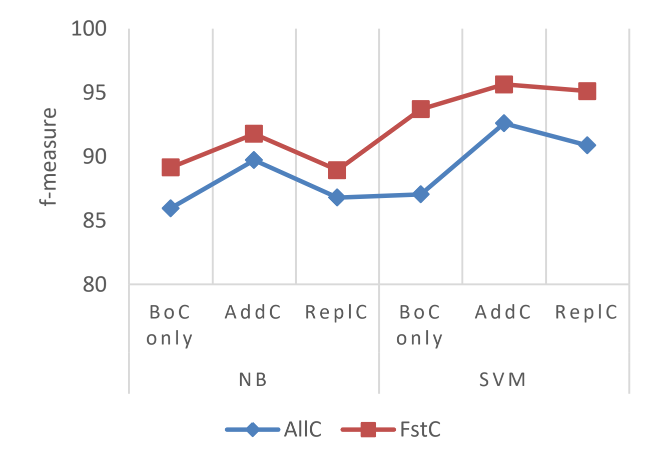


Fig. 4. Comparison of WSD Methods on the Concepts representation with Different Classifier

References

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Found. Trends Inf. Retr., vol. 2, pp. 1-135, 2008.

[2] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, pp. 613-620, 1975.

[3] A. Hotho, S. Staab, and G. Stumme, "Wordnet improves Text Document Clustering," 2003.

[4] Z. Elberrichi, A. Rahmoun, and M. A. Bentaallah, "Using WordNet for Text Categorization," Int. Arab J. Inf. Technol., vol. 5, pp. 16-24, 2008.

[5] Z. Elberrichi and K. Abidi, "Arabic text categorization: a comparative study of different representation modes," Int. Arab J. Inf. Technol., vol. 9, pp. 465-470, 2012.

[6] S. A. Yousif, V. W. Samawi, I. Elkabani, and R. Zantout, "The Effect of Combining Different Semantic Relations on Arabic Text Classification."

[7] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in Contemporary Computing (IC3), 2014 Seventh International Conference on, 2014, pp. 437-442.

[8] A. Balamurali, A. Joshi, and P. Bhattacharyya, "Robust sense-based sentiment classification," in Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2011, pp. 132-138.

[9] Internet World Stats, "Top Ten Internet languages - world Internet statistics," 2016. [Online]. Available: http://www.internetworldstats.com/stats7.htm. Accessed: Oct. 29, 2016

[10] M. K. Saad and W. Ashour, "Arabic morphological tools for text mining," Corpora, vol. 18, p. 19, 2010.

[11] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, p. 14, 2009.

[12] A. Shoukry and A. Rafea, "Preprocessing Egyptian dialect tweets for sentiment mining," in The Fourth Workshop on Computational Approaches to Arabic Script-based Languages, 2012, p. 47.

[13] R. M. Duwairi and I. Qarqaz, "Arabic Sentiment Analysis using Supervised Classification," in Future Internet of Things and Cloud (FiCloud), 2014 International Conference on, 2014, pp. 579-583.

[14] M. Abdul-Mageed, S. Kuebler, and M. Diab, "SAMAR: A system for subjectivity and sentiment analysis of social media Arabic," in 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ICC Jeju, Republic of Korea, 2012.

[15] R. M. Duwairi, "Sentiment analysis for dialectical Arabic," in Information and Communication Systems (ICICS), 2015 6th International Conference on, 2015, pp. 166-170.

[16] M. Hammad and M. Al-awadi, "Sentiment Analysis for Arabic Reviews in Social Networks Using Machine Learning," in Information Technology: New Generations, ed: Springer, 2016, pp. 131-139.

[17] T. Elghazaly, A. Mahmoud, and H. A. Hefny, "Political Sentiment Analysis Using Twitter Data," in Proceedings of the International Conference on Internet of things and Cloud Computing, 2016, p. 11.

[18] R. Baghel and R. Dhir, "A Frequent Concepts Based Document Clustering Algorithm," International Journal of Computer Applications, vol. 4, pp. 6-12, 2010.

[19] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in The Semantic Web–ISWC 2012, ed: Springer, 2012, pp. 508-524.

[20] C. C. Aggarwal and C. Zhai, Mining text data: Springer Science & Business Media, 2012.

[21] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," WSEAS transactions on computers, vol. 4, pp. 966-974, 2005.

[22] C. Quan and F. Ren, "Construction of a blog emotion corpus for Chinese emotional expression analysis," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, 2009, pp. 1446-1454.

[23] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, et al., "Introducing the Arabic wordnet project," in Proceedings of the third international WordNet conference, 2006, pp. 295-300.

[24] F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv., vol. 34, pp. 1-47, 2002.

[25] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," SIGKDD Explor. Newsl., vol. 12, pp. 49-57, 2010.

Logarithmic Spiral-based Construction of RBF Classifiers

Mohamed Wajih Guerfala
Laboratory of research in Automatic Control, ENIT
University of Tunis El Manar, National Engineering School of Tunis
BP 37, Le Belvédère 1002 Tunis, Tunisia

Amel Sifaoui
Laboratory of research in Automatic Control, ENIT
University of Tunis El Manar, National Engineering School of Tunis
BP 37, Le Belvédère 1002 Tunis, Tunisia

Afef Abdelkrim
Laboratory of research in Automatic Control, ENIT
University of Carthage, National Engineering School of Carthage (ENICarthage),
BP 37, Le Belvédère 1002 Tunis, Tunisia

***Abstract*—Clustering process is defined as grouping similar objects together into homogeneous groups or clusters. Objects that belong to one cluster should be very similar to each other, but objects in different clusters will be dissimilar. It aims to simplify the representation of the initial data. The automatic classification recovers all the methods allowing the automatic construction of such groups. This paper describes the design of radial basis function (RBF) neural classifiers using a new algorithm for characterizing the hidden layer structure. This algorithm, called k-means Mahalanobis distance, groups the training data class by class in order to calculate the optimal number of clusters of the hidden layer, using two validity indexes. To initialize the initial clusters of k-means algorithm, the method of logarithmic spiral golden angle has been used. Two real data sets (Iris and Wine) are considered to improve the efficiency of the proposed approach and the obtained results are compared with basic literature classifier**

Keywords—Radial Basis Function neural network; classification; k -means; validity index of Davis Bouldin; Mean Squared Error; Mahalanobis distance; Logarithmic spiral; golden angle; golden ratio

I. INTRODUCTION

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. The problem with Clustering is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters [1].

In recent years, a number of clustering algorithms has been proposed and is available in literature. The radial basis function (RBF) neural network is one of the most used in data classification.

RBF neural networks consist of three layers: an input layer, a hidden layer and an output layer. The input layer corresponds to the input vector feature space and the output layer corresponds to the pattern classes [2]. So the whole architecture is fixed only by determining the hidden layer and the weights between the middle and the output layers [3].

Its training procedure is usually split into two successive steps: training in the hidden layer followed by training in the output layer [4]. First, the centers of the hidden layer (HL) neurons are selected by clustering algorithms such as k-means [5], [6], support vector machine (SVM) [7] or hierarchical clustering [8].Second, the weights connecting the hidden layer with the output layer are determined by Singular Value Decomposition (SVD) or by Least Mean Squared (LMS) algorithms.

One of the used techniques to find the optimal number of this HL is the logarithmic spiral which has seen a significant amount of research on nature-inspired optimization techniques such as neuro-computing in the past 25 years, evolutionary and genetic algorithms, particle swarm optimization. Most recently, a new multipoint meta-heuristics research method has emerged for 2-dimensional continuous optimization problems based on the analogy of spiral phenomena in nature, called 2- dimensional spiral optimization first proposed by Tamura and Yasuda in 2010 [9].

Focused spiral phenomena are approximated to logarithmic spirals, which frequently appear in nature, such as whirling currents, nautilus shells and arms of spiral galaxies. A Two-dimensional spiral optimization uses the feature of Logarithmic Spirals [LS] [9].

In this paper, a new learning algorithm is proposed for the construction of the radial basis function networks solving classification problems. It determines the proper number of hidden neurons automatically and calculates the centers values of radial basis functions. After the selection of the hidden neurons, the widths of nodes are determined by the P-nearest neighbors heuristic, and the weights between the hidden layer and the output layer are calculated by the pseudo-inverse matrix.

The aim of this approach consists in transforming the problem of determining the number of hidden layer neurons to a clustering problem. In order to determine the number of clusters in the data of each class, the k-means algorithm is combined with two different validity indexes (the validity index of Davis Bouldin for the first classifier and Mean Square Error for the second classifier).

In k-means algorithm, the used distance corresponds to the Mahalanobis distance. A solution is also given to overcome the problem of setting the start values for the initial centers needed to start this algorithm using the proposed method "The logarithmic spiral golden angle". Two different real databases are used in order to evaluate the two proposed classifier performances.

Next section presents the problem of the construction of the hidden layer for RBF neural networks. Section 3 describes in detail the logarithmic spiral golden angle. Section 4 is devoted to the elaborated solution to overcome the problem of the k-means algorithm. Section 5 describes the construction of the two proposed RBF classifiers. Experiments and discussions are presented in Section 6, followed by concluding remarks in Section 7.

II. Problem Statement of Use

The construction of the hidden layer of RBF neural networks is by clustering algorithms such as k-means. K-means clustering algorithm is one of the best-known algorithms used in clustering.

However, it still has some problems one of which is in its initialization step which is generally done randomly by users. Another disadvantage of k-means is that it converges to local optimum, depending on its random initialization.

The k-means algorithm classifies objects to a pre-defined number of clusters, which is given by the user (assume k clusters). The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. The Starting points affect the clustering process and results [10], *“Fig. 1”*.

Each boot (initialization) is a different solution (local optimum) which can in some cases be far from the optimal solution (global optimum) [11]. A simple solution to this problem is to run the algorithm several times with different initialization and retain the best combination found.

The use of this solution is limited because of its cost and because of the possibility of finding better results in a single execution [12].

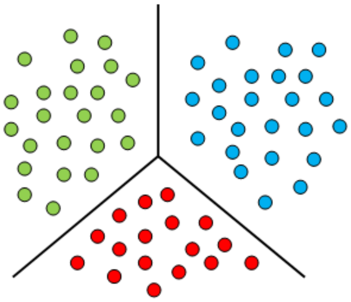


Fig. 1. Clustering on a set of 2D points data, 3 clusters

The main idea of this work is to improve the performance of the k-Means clustering algorithm by fixing its weaknesses. Randomness is one of the techniques wildly used in initializing K-means algorithm that is why it is considered as the main point of weakness that should be solved.

However, because of the sensitivity of k-means to its initial points, two solutions have been proposed to this problem. The first one is to initialize the centers of k-means algorithm using the circle method [13]. The obtained results of this algorithm are the initial centers positions C_k represented by the *“Fig. 2”*.

The second solution is to initialize the centers of k-means algorithm using the logarithmic spiral golden angle in order to improve the clustering performance. In this paper, the second solution will be explained.

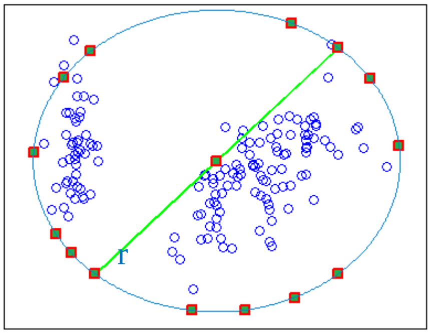


Fig. 2. Tracing of initialization of the k centers maximum on the outline of the circle of radius r and spaced by the angle θ

III. The Logarithmic Spiral Golden Angle

The logarithmic spiral golden angle is a specific case of the logarithmic spiral. It represents a plane curve centered in a starting point and parameterized by the radius r, the angle θ and the Golden Ratio φ.

A. The logarithmic spiral

A Logarithmic Spiral is a plane curve for which the angle between the radius vector and the tangent to the curve is a constant [14]. Such spirals can be approximated mathematically defined by the following equation on the 2-dimensional polar coordinate system (r,θ) as [9]:

$$r = a\,e^{b\theta} \tag{1}$$

Where a and b are positive real with $a>0$ and $b\neq 0$.

Equation (1) can be transformed into Cartesian coordinates as follows:

$$\begin{cases} x(\theta)=r(\theta)\ \cos(\theta)=ae^{b\theta}\cos(\theta) \\ y(\theta)=r(\theta)\ \sin(\theta)=ae^{b\theta}\sin(\theta) \end{cases} \tag{2}$$

In this work, the factor b of the logarithmic spiral has been set to zero ($b=0$), it goes back to simplify the polar radius as follows: $r=ae^{b\theta}=ae^{0\times\theta}\Rightarrow r=a$

The following equation of the logarithmic spiral golden angle is obtained:

$$\begin{cases} x(\theta) = r\ \cos(\theta) = a\ \cos(\theta) \\ y(\theta) = r\ \sin(\theta) = a\ \sin(\theta) \end{cases} \tag{3}$$

B. The Golden ratio

The irrational number, golden ratio is also known as golden section by the ancient Greeks, golden proportion, divine proportion or golden number [15].

The golden ratio φ, has many properties which people are eager to know. It is a number that is equal to the reciprocal of its own with the addition of 1: $\varphi = \frac{1}{\varphi} + 1$.

Likewise, the ratio of any two consecutive Fibonacci numbers converges to give approximates of 1.618, or its inverse, 0.618. This shows the relationship between Fibonacci numbers and golden ratio [16].

If the possibility of dividing a line in such a way that the ratio of the whole length to the length of the longer segment happens to be equal to the ratio of the length of the longer segment to the length of the shorter segment, then it could be said that the ratio is a golden ratio [15], *"Fig. 3"*.

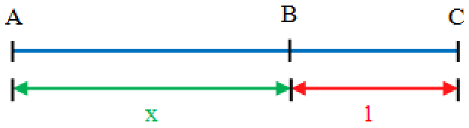


Fig. 3. Dividing of a whole length AC into two segments AB and BC

This gives mean ratio if $\frac{AB}{BC} = \frac{AC}{AB}$. If the value of AB is set to be x, and use 1 to represent the length of BC, then $\frac{x}{1} = \frac{1+x}{x}$ is obtained. Then the irrational number is the only positive solution of the equation $x^2 - x - 1 = 0$, so $x = \frac{1+\sqrt{5}}{2}$

Its value is: $\varphi = \frac{1+\sqrt{5}}{2} \approx 1{,}6180339887$ .Where the Greek letter phi (φ) represents the golden ratio.

C. The Golden angle

In geometry, the golden angle is created by dividing the circumference of a circle c in two sections, a longer arc of length a and a smaller arc of length b such that: $c = a + b$ and $\varphi = \frac{a+b}{a} = \frac{c}{a} = \frac{a}{b}$, *"Fig. 4"*.

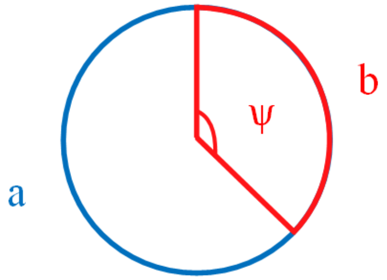


Fig. 4. Golden angle measurement

The angle formed by the arc b of circle c is called the golden angle ψ. It derives from the golden ratio φ.

$$\psi = 2\pi - \left(\frac{2\pi}{\varphi}\right) = 2\pi\frac{(\varphi - 1)}{\varphi} \approx 2{,}391 \text{ rad} \approx 137{,}5°$$

IV. Proposed Initialization of the K-Means Algorithm with Logarithmic Spiral Golden Angle

The k-means algorithm aims to minimize the distance between the object and the center of its group. In this section, the k-means algorithm based on the Mahalanobis distance and its proposition for initialization of the centers are presented.

A. The k-means algorithm Mahalanobis distance specifications

There are different types of distances such as: Minkowski distance, the average, the family of metrics, Euclidean Weighted and the Euclidean distance which is the most used, e.g. applied in the RBF Networks. [17]

Moreover, the Mahalanobis distance is a distance measure and its utility is a way to determine the similarity between two multidimensional random variables. It differs from Euclidean distance, because the Mahalanobis distance takes into account the correlation between random variables, [17]. The Mahalanobis distance is defined by:

$$d(x,y) = \sqrt{(x-y) \times Cov(X)^{-1} \times (x-y)^T} \tag{4}$$

With $Cov(X)$ is the covariance matrix.

If the elements x and y are independent, the covariance matrix is the identity and the Mahalanobis distance is equal to the Euclidean distance. The proposed algorithm based on the combination of Mahalanobis distance with k-means is described by the following steps:

Algorithm: Function Kmeans_distance_Mahalanobis (KDM)

```
Begin
Input:  - The database $X=\{x_1,x_2,.....,x_N\}\in R^d$ .
- The position of each center $C=\{c_1,c_2,...,c_k\}\in R^d$ .
Output: - The new position of each center $C^*=\{c_1^*,...,c_k^*\}\in R^d$
Step 1:
- Determine the size $N$ of the data base of $X$ .
- Determine the number $k$ of centers to be used in the observation space $C$ .
- Initialize the vector of the new positions of the centers $C^*$ to zero.
Step 2:  - Determine the covariance matrix $Cov(X)$ with the following equation:
      $Cov(X)=\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij}-\bar{X}_j)(X_{ij}-\bar{X}_j)^T$
with $X_{ij}\in X$ ; $i=1,...,n$ and $j=1,...,p$ .
Where $\bar{X}_j=\sum_{i=1}^{n}X_{ij}$ with $\bar{X}_j$ : arithmetic averages.
While: The new centers do not have, a significant displacement Do:
    Step 3: - Assign each observation (dot) group nearest center $c_j : x_l\in c_j$ according to the Mahalanobis distance formula:
          $d(i,j)=\sqrt{(x_i-c_j)\times Cov(X)^{-1}\times(x_i-c_j)^T}$
    with $l=1,...,N$ and $j=1,...,k$ .
    Step 4: - Recalculate the position of each new center:
          $c^*_j=\frac{1}{N_j}\sum_{x_l\in c_j}^{N}x_l$
    with $N_j$ = the set of points belonging to the center $c_j$ and $j=1,...,k$ .
End While
End
```

To increase the performance of the k-means algorithm Mahalanobis distance, a solution is proposed to initialize the k-means algorithm using the logarithmic spiral golden angle parameterized by the radius r, the angle θ and the Golden ratio φ and the Golden angle ψ [18]. This solution is divided into several steps:

The first step is to calculate the maximum distance between two individual points (a,b) belonging to the database, then to define the middle ground G between these two individual points and determine the radius $R=\overline{Gb}$ *"Fig. 5"*.

The second step is to calculate the golden number φ by applying the following formula: $\varphi=\frac{1+\sqrt{5}}{2}\approx 1{,}6180339887$

The third is to initialize the values of the logarithmic spiral golden angle on the polar coordinate system (r,θ): the radius $r=a=a_0=0$ and the angle $\theta=\theta_0=0$ .The angle θ increases by the factor $d\theta=\psi=2\pi\frac{(\varphi-1)}{\varphi}$ and the radius r increases by the factor $da=\frac{R}{k_{\max}}=\frac{\overline{Gb}}{k_{\max}}$.

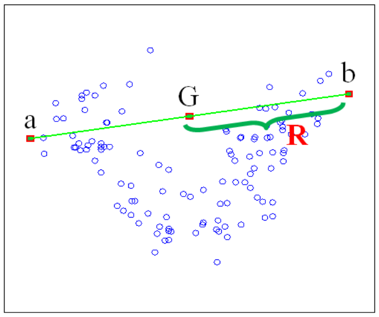


Fig. 5. Tracing of the two most distant individuals $(a,\ b)$ and their medium G

To determine $k_{\max}$, the suggestion of Bezdek was adopted [19] as follows : $k_{\max}=\sqrt{N}$,(N is the size of the database) .

The forth step is to determine the positions of the centers of the logarithmic spiral golden angle with center G and radius $r=a$ *"Fig. 6"*. The calculation of the center position C_k is performed by applying the following formula:

$$\begin{cases} C_{kx}=G_x+a\times\cos(\theta) \\ C_{ky}=G_y+a\times\sin(\theta) \end{cases} \quad (5)$$

With $\theta=\theta+d\theta$; $a=a+da$ and $k=1,...,k_{\max}$.

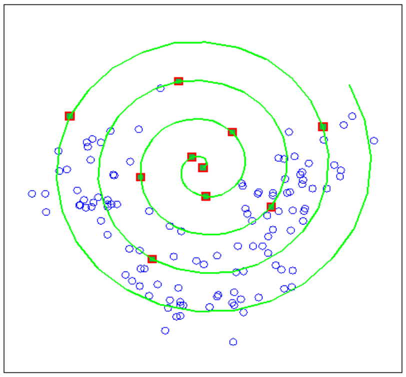


Fig. 6. Tracing of the initialization of the k centers maximum on the outline of the logarithmic spiral golden angle

Thereafter, the positions of these $k_{\max}$ centers in the variable $C_k=\{c_1,...,c_{k\max}\}$ will be saved.

The basic principle of the adopted strategy is summarized in the following algorithm:

Algorithm: Init_Centres_Kmeans_ Logarithmic_Spiral

```
Begin
Input:   - The database $X=\{x_1,x_2,.....,x_N\}\in R^d$ .
 - The maximum number of centroids $k_{\max}$ .
Output: - The position of each center $C=\{c_1,..,c_k,...,c_{k\max}\}\in R^d$
Step 1:   - Calculate the maximum distance $D$ between two points belonging to the base $X$ .
 - Calculate the center $G$ of $D$ and the radius $R=\overline{Gb}$ .
Step 2:   - Calculate the golden ratio $\varphi$ by applying the following formula: $\varphi=\frac{1+\sqrt{5}}{2}$
Step 3:   - Initialize the values of the logarithmic spiral golden angle on the polar coordinate system $(r,\theta)$: the radius $r=a=a_0=0$ and the angle $\theta=\theta_0=0$ .
    - Fix the increment of the angle $\theta$ by the factor $d\theta=\psi=2\pi\frac{(\varphi-1)}{\varphi}$
    - Fix the increment of the radius $r$ by the factor $da=\frac{R}{k_{\max}}=\frac{\overline{Gb}}{k_{\max}}$
Step 4: - Determine the positions of the centers belonging to the logarithmic spiral golden angle with center $G$ and radius $r=a$ according to the following formula:
            $\begin{cases} C_{kx}=G_x+a\times\cos(\theta) \\ C_{ky}=G_y+a\times\sin(\theta) \end{cases}$
 With $\theta=\theta+d\theta$ ; $a=a+da$ and $k=1,...,k_{\max}$ .
Step 5: - Save the positions of the centers found in $C=\{c_1,..,c_k,...,c_{k\max}\}\in R^d$ .
End
```

B. Evaluation Measures

Using an unsupervised clustering algorithm, such as k-means algorithm, requires the determination of the number k of groups leading to the execution of the algorithm repeatedly for different values of this parameter.

For optimal number of groups, a criterion should be used to evaluate the result of the algorithm. This criterion is known as the validity index [1], [20]-[22], [15, 16, 17, and 18] name based on the notions of compactness and separation.

In literature, there are a lot of validity indexes, most of them are based on the notions of compactness within different groups and the separability between these different groups. In this article, the Davies-Bouldin index and the Mean Squared Error will be used as two validity indexes of neural classifiers.

C. Davies-Bouldin Index

This index takes into account both of the compactness and the separability of groups [23]. Its value is much lower than the groups are compact and well separated. It promotes hyperspherical groups and is, therefore, particularly well-suited for a use with the k-means algorithm. The I_{DB} index is defined by the following expression:

$$I_{DB}=\frac{1}{K}\sum_{i=1}^{k}\max_{i\neq j}\frac{\{d_c(c_i)+d_c(c_j)\}}{D_{cc}(c_i,c_j)} \tag{6}$$

Where $d_c(c_i)$ is the average distance between an object and its group c_i following the center and $D_{cc}(c_i,c_j)$ is the distance between the centers of groups c_i and c_j with:

$$d_c(c_i)=\frac{1}{N_l}\sum_{l=1}^{N_l}\|x_l-c_j\| \tag{7}$$

$$D_{cc}(c_i,c_j)=\|c_i-c_j\| \tag{8}$$

D. Mean Squared Error

The Mean Squared Error is frequently used to assess the risk of an estimator. It is also useful to relay the concepts of bias, precision, and accuracy in statistical estimation. In this work, the MSE was used for groups' compactness measure for each centroid [24]; it is notably equivalent to the Euclidean function of the k-means algorithm:

$$E=\sum_{l=1}^{N}\sum_{j=1}^{k}\delta_{il}\left\|x_l-c_j\right\|^2 \tag{9}$$

$$\text{With: } \delta_{il}\begin{cases}1 \text{ if } x_l\in c_i \\ 0 \text{ else}\end{cases}$$

V. New Algorithms for the Construction of the Hidden Layer of the RBF Classifier

Two new algorithms were used to characterize the hidden layer classifier i.e. to determine the number of centers of different Gaussian and the value of each center.

In what follows, the principle of the proposed classifiers is presented. It is also explained how the two validity indexes (IDB and MSE) are combined with the k-means algorithm Mahalanobis distance to determine automatically the number k of groups.

However, it is necessary to fix a maximum number of centroids $k_{\max}$. The $k_{\max}$ value can be defined by the user if he/she knows the structure of his database. Given that it is not always the case, the Bezdek [19] suggestion is adopted, so the $k_{\max}=\sqrt{N}$ (N is the size of the database) is chosen.

Applying these algorithms to all classes and summing the number of the obtained groups, the number of neurons in the hidden layer is determined. A neuron is then assigned to each group. For this RBF classifier, the database was partitioned $X=\{x_1,x_2,...,x_N\}\in R^d$ in individual blocks according to the number of output classes $\Omega_j = 1,2,...,m$ .The database $X_d = \left(X_{\Omega 1} \quad X_{\Omega 2} \quad . \quad . \quad . \quad X_{\Omega\, m}\right)^T$ is obtained.

Then, we apply the Principal Component Analysis (PCA) to the data base X_d in order to reduce it to a new basis two-dimension $X^*{}_d = \left\{x_1, x_2\right\} \in R^2$.

Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. K-means clustering is a commonly used data clustering for performing unsupervised learning tasks [25].

The PCA is based on the calculation of averages, variances and correlation coefficients. The main basis of dimension reduction is that PCA picks up the dimensions with the largest variances.

The PCA allows, in the same time, a reduction of data and an easier interpretation in the treated domain, as the new dimensions are often very significant [26]. In this case, the two largest variances were chosen to represent the new database.

The next step is to determine the number of centers and the center position of each class $C_{\Omega 1} = \left\{c_1,...,c_k\right\} \in R^d$ through the classifier based on the k-means algorithm with Mahalanobis distance.

The centers of each class $\Omega_j = 1,...,m$ found in the $C = \begin{pmatrix} C_{\Omega 1} \\ . \\ . \\ C_{\Omega m} \end{pmatrix}$ matrix are grouped and the k-means algorithm Mahalanobis distance is applied to the new positions of the centers $C^* = \left\{c_1^*, c_2^*, ..., c_K^*\right\} \in R^d$.

To complete the construction of the hidden layer classifier, there is a second parameter to consider in the neurons, which is the width factor σ_j for each centroid c_j ($j = 1,...,k$). This factor is calculated using the following formula:

$$\sigma_j = \frac{1}{N_a\sqrt{8}}\sum_{i=1}^{Na}\left\|x_i - c_j\right\| \qquad (10)$$

Witch N_a represents the training data

A. Construction of the RBF classifier KMD-LS-IDB

The proposed algorithm (KMD-LS-IDB) based on the k-means algorithm with Mahalanobis distance combined with the Davies-Bouldin validity index. This classifier determining the number and the centers values of the hidden layer for each class of the database is described below:

```
Begin
Input:
- The block database $X_{\Omega j} = \left\{X_{\Omega 1},...,X_{\Omega m}\right\} \in R^d$ of one class of data base $X_d$ , taking the case of the block. (The same approach for other classes).
Output: - The position of each center $C_{\Omega 1} = \left\{c_1,...,c_k\right\} \in R^d$ .
Step 1: - Determine the size $n$ and the number of characters (the attributes) $p$ of the data base $X_{\Omega 1}$ .
Step 2: - Initialize the minimum number of centroids $K_{\min} = 2$ and then look for the maximum number of centroids by $K_{\max} = \sqrt{n}$ .
  - Initialize the variables $a = k = K_{\min}$ and $d = 1$.
Step 3: - Apply Algorithm Init_Centres_Kmeans_Logarithmic_Spiral which initializes the centers for kmeans algorithm of the data base $X_{\Omega 1}$ .
Step 4: - Repeat the following steps until $k = K_{\max}$
  Step 4.1: -If $k \prec K_{\max} + 1$ Then:
      - Take the following centers positions $C = \left\{c_1,...,c_a\right\} \in R^d$ .
      - Deduce the number of the centers $k$ .
    End If
  Step 4.2: - Apply k-means algorithm with Mahalanobis distance to determine the new positions of the $C^* = \left\{c_1^*,...,c_a^*\right\} \in R^d$ centers.
  Step 4.3: - Calculate the compactness and separability of groups with the Davies- Bouldin Index: $I_{DB}$ .
  Step 4.4: - Save $I_{DB}$ variable in the table called $Tab\_IDB$ .
  Step 4.5: - Increment variables $a = a + 1$, $d = d + 1$ .
Step 5: - Determine $\beta$ : the $I_{DB}$ lowest index of the table $Tab\_IDB$ .
  - Take the following centers positions $C_{\Omega 1} = \left\{c_1,...,c_{\beta+1}\right\} \in R^d$ as the optimal number required classifying $\Omega_1$ class centers.
End
```

B. Construction of RBF classifier KMD-LS-MSE

The proposed algorithm (KMD-LS-MSE) based on the k-means algorithm with Mahalanobis distance combined with the Mean Squared Error validity index. This classifier determining the number and the centers values of the hidden layer for each class of the database is described below:

```
Begin
Input:
- The block database $X_{\Omega j} = \{X_{\Omega 1},...,X_{\Omega m}\} \in R^d$ of one class of data base $X_d$, taking the case of the block. (The same approach for other classes).
Output: - the position of each center $C_{\Omega 1} = \{c_1,...,c_k\} \in R^d$.
Step 1: - Determine the size $n$ and the number of characters (the attributes) $p$ of the data base $X_{\Omega 1}$.
Step 2: - Initialize the minimum number of centroids $K_{\min} = 2$ and then look for the maximum number of centroids by $K_{\max} = \sqrt{n}$.
- Initialize the variables $a = k = K_{\min}$ and $d = 1$.
Step 3: - Apply Algorithm Init_Centres_Kmeans_Logarithmic_Spiral which initializes the centers for kmeans algorithm of the data base $X_{\Omega 1}$.
Step 4: - Repeat the following steps until $k = K_{\max}$
Step 4.1: -If $k \prec K_{\max} + 1$ Then:
    - Take the following centers positions $C = \{c_1,...,c_a\} \in R^d$.
    - Deduce the number of centers $k$.
          End If
Step 4.2: - Apply k-means algorithm with Mahalanobis distance to determine the new positions of the $C^* = \{c_1^*,...,c_a^*\} \in R^d$ centers.
Step 4.3: - Calculate the compactness and separability of groups with the Mean Squared Error Index: $M_{SE}$.
Step 4.4: -Save $M_{SE}$ variable in the table called $Tab\_MSE$.
Step 4.5: - Increment variables $a = a+1$, $d = d+1$.
Step 5: - Determine $\beta$: the $M_{SE}$ lowest index of the table $Tab\_MSE$.
- Take the following centers positions $C_{\Omega 1} = \{c_1,...,c_{\beta+1}\} \in R^d$ as the optimal number required classifying $\Omega_1$ class centers.
End
```

C. Calculation of synaptic weight

After determining the parameters of the proposed classifier hidden layer, the learning is finished by the calculation of the synaptic weight w_{ij}, connecting the hidden layer neurons to those of the output layer.

The linearity property of the outputs $y_j(x_l)$ of the network is used**.** The expression of each of the m outputs is written as:

$$y_j(x_l) = h_j(x_l) \times w_{ij} \tag{11}$$

The global output of the network is written as follows:

$$Y = H \times W \tag{12}$$

The objective is to determine the matrix W that minimizes an error function, chosen as the square of the sum of classification errors.

The weight of the output layer can be calculated by the following matrix equation:

$$\underbrace{\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots\cdots & \varphi_{1M} \\ \varphi_{21} & \varphi_{22} & \cdots\cdots & \varphi_{2M} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ \varphi_{N1} & \varphi_{N2} & \cdots\cdots & \varphi_{NM} \end{bmatrix}}_{H} \times \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_M \end{bmatrix}}_{W} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix}}_{Y} \tag{13}$$

With

H : The matrix of Radial Basis Functions

Y : The matrix of the output layer.

W : The weight matrix of centroids.

φ_{ij} : The Gaussian widths of the matrix H .

The above equation is giving by:

$$H \times W = Y \Rightarrow W = H^{-1} \times Y \tag{14}$$

Given that the H matrix is rarely square, the pseudo-inverse of the matrix H is applied according to equation (15):

$$W = \left[H^T \times H\right]^{-1} \times H^T \times Y \tag{15}$$

VI. Evaluation oF RBF Classifiers

The purpose of this section is to evaluate the performance and efficiency of the proposed RBF neural classifiers. The performance of these RBF neural networks classifiers is tested with two different databases: Iris and Wine among the different data sets available from the machine learning community by the University of California at Irvine (UCI) [27].

The first test is carried out with the Wine database which includes the results of a chemical analysis of types of wine produced in different regions of Italy from different grapes varieties. The concentration of 13 components are indicated for each of the 178 types of wine (patterns) which are analyzed and divided into three classes (59 in Class 1, 71 in Class 2 and 48 in Class 3).

The second test is done with Iris database which is one of the most popular data set to examine the performance of novel methods in pattern recognition and machine learning. It is composed of three classes (i.e., iris Setosa, iris Versicolor and iris Virginica) each having 50 patterns with four features.

To evaluate the proposed classifiers performances, the holdout method is used. It consists on dividing the initial data into two independent sets: one for training and the other for testing the classifier performances.

The results given by the RBF classifier built with this algorithm are compared with those obtained with other neural classifiers: the Learning Vector Quantization (LVQ) classifier proposed by Kohonen, the RBF neural network classifier for which the hidden layer is obtained using adaptive Pattern Classifier (APCIII) [28], the Multi-Layer Perceptrons classifier (MLP) and with a reference: the K nearest Neighbor (KNN).

The present comparative results of different classifiers over Wine and Iris are illustrated in *Table I* and *Table II*.

TABLE I. RESULTS OF THE RECOGNITION RATE OVER WINE DATABASE

Classification algorithms	**Database: Wine**
KMD-LS-IDB	98,88 %
KMD-LS- MSE	95.55 %
LVQ	66,14 %
APCIII	67,04 %
MLP	73,80 %
KNN	70,45 %

TABLE II. RESULTS OF THE RECOGNITION RATE OVER IRIS DATABASE

Classification algorithms	**Database: Iris**
KMD-LS-IDB	93.46 %
KMD -LS-MSE	93.46 %
LVQ	94,00 %
APCIII	93,33%
MLP	96,66 %
KNN	96,70 %

Considering Wine database, the best recognition rate is obtained by the KMD-LS-IDB proposed classifier and then the proposed classifier KMD-LS-MSE. For Iris database, the best recognition rate is given for the KNN classifier; however, the difference with the two proposed classifiers is not important.

Then, the proposed algorithms give good results in terms of recognition rate but the most powerful of them is the classifier KMD-LS-IDB.

VII. CONCLUSION

In this paper, new RBF neural networks classifiers has been designed to classify database. The proposed algorithms aim to deduce the centers of the hidden layer neurons and to calculate the number of the neurons in particular.

The basic idea of this approach is to select the training data from the database class by class and to decide about the optimal number of neurons in each class by using two different validity indexes (the validity index *IDB* and the MSE). This number is integrated in the k-means algorithm with the Mahalanobis distance.

A solution was also proposed to overcome the problem of initialization of centers necessary for the start of the K-means algorithm using the method of the logarithmic spiral golden angle.

The obtained classifiers results are satisfactory in comparison with other considered classifiers in literature for two real databases (Iris and Wine).

REFERENCES

[1] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” J. Intell. Inf. Syst., vol. 17, no. 2–3, pp. 107–145, 2001.

[2] J. K. Sing, D. K. Basu, M. Nasipuri, and M. Kundu, “Improved k-means algorithm in the design of RBF neural networks,” Conf. Converg. Technol. Asia-Pacific Reg. TENCON, pp. 841–845, 2003.

[3] S. Song, Z. Yu, and X. Chen, “A novel radial basis function neural network for approximation,” Int. J. Inf. …, vol. 11, no. 9, pp. 46–53, 2005.

[4] B. Mhamdi, T. Aguili, and K. Grayaa, “Radial Basis Function Neural Network to Shape Reconstruction of Conducting Objects,” 2012 6th Int. Conf. Sci. Electron. Technol. Inf. Telecommun. SETIT 2012, vol. 1, no. 2, pp. 628–633, 2012.

[5] J. Moody and C. J. Darken, “Fast Learning in Networks of Locally-Tuned Processing Units,” Neural Comput., vol. 1, pp. 281–294, 1989.

[6] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.

[7] M. Vogt, “Combination of Radial Basis Function Neural Networks with Optimized Learning Vector Quantization,” Proc. ICNN’93, Int. Conf. Neural Networks, vol. III, pp. 1841–1846, 1993.

[8] M. Kubat, “Decision trees can initialize radial-basis function networks,” IEEE Trans. Neural Networks, vol. 9, pp. 813–821, 1998.

[9] N. Siddique and H. Adeli, “Spiral Dynamics Algorithm,” vol. 23, no. 6, pp. 1–24, 2014.

[10] B. Al-shboul and S. Myaeng, “Initializing K-Means using Genetic Algorithms,” Int. J. Comput. Electr. Autom. Control Inf. Eng., vol. 3, p. 6, 2009.

[11] Sifaoui, A. Abdelkrim, and M. Benrejeb, “On the Use of Neural Network as a Universal Approximator,” Int. J. Sci. Tech. Autom. Control Comput., vol. 2, no. July, pp. 386–399, 2008.

[12] Sifaoui, A. Abdelkrim, and M. Benrejeb, “On New RBF Neural Network Construction Algorithm for Classification,” SIC, vol. 18, no. 2, pp. 103–110, 2009.

[13] M. W. Guerfala, A. Sifaoui, and A. Abdelkrim, “RBF Neural Network Construction Algorithm for Classification based on Mahalanobis distance,” in ACECS’15, 2015, no. 1, pp. 1–5.

[14] R. H. Bacon, “Logarithmic Spiral: An Ideal Trajectory for the Interplanetary Vehicle with Engines of Low Sustained Thrust,” Am. J. Phys., vol. 27, no. 3, p. 164, 1959.

[15] G. Markowsky, “Misconceptions about the Golden Ratio,” Coll. Math. J., vol. 23, no. 1, pp. 2–19, 1992.

[16] L. D. G. Sigalotti and A. Mejias, “The golden ratio in special relativity,” Chaos, Solitons & Fractals, vol. 30, no. 3, pp. 521–524, 2006.

[17] R. J. Praga-Alejo, L. M. Torres-Treviño, D. S. González-González, J. Acevedo-Dávila, and F. Cepeda-Rodríguez, “Analysis and evaluation in a welding process applying a Redesigned Radial Basis Function,” Expert Syst. Appl., vol. 39, no. 10, pp. 9669–9675, 2012.

[18] M. W. Guerfala, A. Sifaoui, and A. Abdelkrim, “Construction of an RBF Classifier Based on Logarithmic Spiral,” in ACECS’16, 2016, pp. 1–6.

[19] J. C. Bezdek, R. J. Hathaway, M. J. Sabin, and W. T. Tucker, “Convergence Theory for Fuzzy C-Means: Counterexamples and Repairs,” Syst. Man Cybern. IEEE Trans., vol. 17, no. 5, pp. 873–877, 1987.

[20] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “Clustering algorithms and validity measures,” Proc. Thirteen. Int. Conf. Sci. Stat. Database Manag. SSDBM 2001, pp. 3–22, 2001.

[21] M. Halkidi and M. Vazirgiannis, “Clustering Validity Assessment : Finding the optimal partitioning of a data set,” in In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 2001, pp. 187–194.

[22] M. Sassi, A. G. Touzi, and H. Ounelli, “Using Gaussians Functions to Determine Representative Clustering Prototypes,” pp. 1–5, 2006.

[23] D. L. Davies and D. W. Bouldin, “A cluster separation measure.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 1, no. 2, pp. 224–227, 1979.

[24] H. Rady, “Reyni ’ s Entropy and Mean Square Error for Improving the Convergence of Multilayer Backprobagation Neural Networks : A Comparative Study,” no. October, 2011.

[25] Ding and X. He, “K-means Clustering via Principal Component Analysis,” Int. Conf. Mach. Learn., 2004.

[26] S. Jaffali and S. Jamoussi, “Principal component analysis neural network for textual document categorization and dimension reduction,” 2012 6th Int. Conf. Sci. Electron. Technol. Inf. Telecommun. SETIT 2012, vol. 1, no. 2, pp. 835–839, 2012.

[27] L. Blake and C. J. Merz, “UCI Repository of machine learning databases,” University of California. p. http://archive.ics.uci.edu/ml/, 1998.

[28] Y. Hwang and S. Bang, “An Efficient Method to Construct a Radial Basis Function Neural Network Classifier,” Neural Networks, 1997.

Priority Task Scheduling Strategy for Heterogeneous Multi-Datacenters in Cloud Computing

Naoufal Er-raji
Department Of Mathematics and Informatics
Laboratory of Modeling and Information Technology
Faculty of sciences Ben M'SIK University Hassan II
Casablanca, Morocco

Faouzia Benabbou
Department Of Mathematics and Informatics
Laboratory of Modeling and Information Technology
Faculty of sciences Ben M'SIK University Hassan II
Casablanca, Morocco

***Abstract*—With the rapid development in science and technology, cloud computing has emerged to be widely adopted in several IT (Information Technology) areas. It allows for the companies as well as researchers to use the computing resources as a service over a network as internet without owning the infrastructure. However, Due to increasing demand of cloud computing, the growing number of tasks affects the system load and performance. Scheduling of multitasks with respect SLA (Service Level Agreement) can face serious challenges. In order to overcome this problem as well as provide better quality of service, the tasks have to be scheduled in optimal way. In this paper, we address the problem of the priority task scheduling through proposing a global strategy over distributed data-center in cloud computing basing on three parameters: tasks deadline, task age and the task length.**

***Keywords*—age; cloud computing; cluster; data-center; deadline; length; node; SLA; priority task scheduling**

I. Introduction

In order to satisfy the rising demand of computing resources, most of the IT industry's and companies start using the cloud computing. This technology is a large-scale distributed computing paradigm that consists of a mix of technologies as virtualization, SOA (Service Oriented Architecture) ... for providing shared pool of computing resources dynamically to the CSC (Cloud Service Consumers) through connecting tens of thousands of servers over a network as internet. Moreover, it offers an easy access to those resources anywhere and at any time [1] [2] [3]. Due to the increasing demand of cloud resources and the limited resources of CSP (Cloud Service Providers), this raises several challenges, and task scheduling is one of them. The task scheduling, is a process of choosing the best suitable available resources for execution the tasks or to allocate computer machines to tasks in such a manner that the completion time is minimized as possible [4] [5].

In this paper, we have focused on the priority tasks because they have a big impact on the services providing and a longer delay can make a violating in the SLA contract. Therefore, we propose a global priority task scheduling strategy that's based on three parameters: task deadline, task length and task age. The main purpose of this strategy is to improve the priority task scheduling, therefore, respecting the deadlines that are documented in the SLA contract. The rest of paper is organized as follow, Section 2 describes the Proposed Model Architecture Cloud Computing, section 3 presents related works, and section 4 explains the priority task scheduling strategy. At last we make a discussion and conclusion of the paper.

II. Proposed Model Architecture Cloud Computing

In the cloud computing, the scheduling can affect the performance in a right or a wrong way. So according to the scheduling process, tasks could be processed quickly or could remain too long time at queue.

The scheduling process can be done at different levels and at different components depending on the nature of the strategy followed by the CSP. Thus, the cloud computing need to be evolved by a good task scheduling strategy and a global view of all processes is needed. This section describes the suggested model. We consider that each CSP has a multi-cloud data-center, each data-center has several clusters and each cluster consists of numerous servers and each server runs numerous VMs (Virtual Machines).

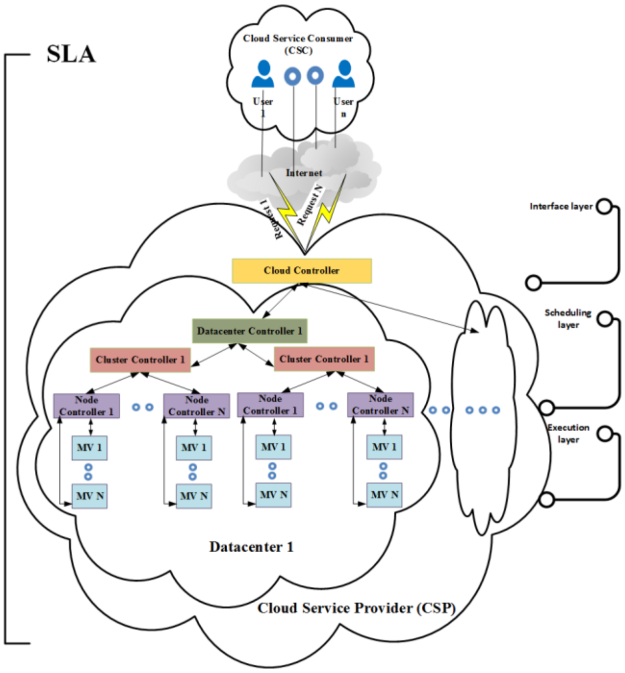


Fig. 1. Proposed Model Architecture Cloud Computing

As shown in the figure one: the proposed Model Architecture Cloud Computing is composed of three layers and six components [6], the characteristics of each one are summarized in the following section. We have the layers:

- **Interface layer**: the responsible for the communication between the CSC and CSP.
- **Scheduling layer**: here where all algorithms and mechanisms worked to schedule the tasks to the suitable resources.
- **Execution layer**: here where the entire task will be executed.

The components are:

- **Users**: they place theirs tasks with the help of cloud controller through their browsers or applications.
- **Cloud Controller**: receive the tasks from the users; keep track of the data-centers and their performance and assign tasks to the suitable data-center.
- **Datacenter Controller**: receive the tasks from the cloud controller, keep track of the clusters and their performance and assign tasks to the suitable cluster according to the cloud controller direction.
- **Cluster Controller**: receive the tasks from the Datacenter controller, keep track of the Nodes and their performance and assign tasks to the suitable Node according to the cloud controller direction.
- **Node Controller:** receive the tasks from the cluster controller, keep track of the VMs and their performance and assign tasks to the suitable VM according to the cloud controller direction.
- **VM**: execute the tasks according to the cloud controller direction.

The CSC submits their tasks having different length, and different deadline (QoS (Quality of Service) requirement), from anywhere at any time, so as to be executed as soon as possible under the SLA constrain. In the other side, the CSP have several data-centers geographically distributed, which consist of numerous servers and each server runs numerous VMs and finally each VM has different capability to execute different tasks. However, with the growing amount of the CSC tasks, with the presence of different priorities tasks, arise several difficulties either for making the best choice of resources or in which order the tasks have to be assigned.

III. Related Works

In the literature, there is a vast number of propositions for improving task scheduling in cloud computing. One of those methods is task scheduling based on priority. Here, some contributions that are based on priority for their propositions:

Atul Vikas Lakraa and Dharmendra Kumar Yadav [7] proposed an algorithm, which assigns priority to different tasks according to the QoS of request task. High QoS task assigned with low QoS value and the low QoS task assigned with high QoS value. Hence, the task with lower QoS value is a high priority and the task with high QoS value is a low priority. And for the resource allocation, the VMs are selecting according to their MIPS (Million Instructions Per Second) Values such that the one having highest MIPS has the highest ability to be assigned.

Samia Ijaz, Ehsan Ullah Munir, Waqas Anwar, and Wasif Nasir [8] proposed a strategy, which assigns priority to different tasks according to the ALST (Absolute Latest Start Time) such that the one having minimal ALST among all the tasks has higher priority. For resource allocation, the VMs are selecting according to EST (Earliest Start Time) and ECT (Earliest Completion Time) which on each VM is reckoned using three mathematical equations and the VM that gives the least ECT for a task is selected.

In Papers [9] [10] the proposed algorithms assign priority to different tasks according to task size such that one having highest size has highest priority. For the resource allocation, the VMs are selecting according to their MIPS values such that the one having highest MIPS has the highest priority to be selected.

Deepika Saxena, R.K. Chauhan and Ramesh Kait [11] proposed an algorithm, which Classify and group all tasks according to their deadline and cost constraints, and assign them to different priority queue (high, mid, low). And for resources allocation the approach is based on greedy resource (VM) allocation for selecting the resources which means that scheduler select VM with minimum turnaround time for each individual task.

Aditi Sharma and Shivi Sharma [12] proposed a technique, which is based on credit system. Each task is assigned a unique credit based upon three parameters namely: Task Length Credit, Task priority credit, Task deadline credit. Based on these credits the tasks are assigned to VM.

Shamsollah Ghanbari and Mohamed Othman [13] propose an algorithm, which is based on the theory of a mathematical model named AHP (Analytical Hierarchy Process) for calculating the priority. It is a MCDM (Multi Criteria Decision Making) and MCDM (Multi Attribute Decision Making) model. The proposed architecture is consisted of three levels namely: objective level, attributes level and alternatives level.

Pankajdeep Kaur and Parampreet Singh [14] proposed an algorithm, which assigns priority to different tasks according to specific attribute: User Level, Task urgency, Task Load and Time queuing up. Then, tasks are arranged in a sorted order by considering the calculated priority. Thus, the task with higher priority scheduled first.

In the literature, there are many contributions. Which make a comparison of approaches either for priority based or other parameters based [15][16][17][18][19][20] [21].

In this section, we propose a comparison of the above propositions. The criteria of comparison are:

- Factor considered for task priority.
- Factor considered for VM priority.
- Centralized Scheduling or distributed scheduling.
- Resources Load.

TABLE I. ALGORITHMS COMPARISON TABLE

Propositions	Factor considered for task priority	Factor considered for VM priority	Centralized Scheduling	Distributed Scheduling	Resources Load
[7]	QoS	MIPS	Yes	No	No
[8]	ALST	EST, ECT	Yes	No	No
[9]	Task SIZE	MIPS	Yes	No	Yes
[10]	Task SIZE	MIPS	Yes	No	No
[11]	DEADLINE, COST	GREEDY (minimum turnaround time)	Yes	No	No
[12]	CREDITS (Task Length Credit, Task Priority Credit, Task Deadline Credit)	No	Yes	No	No
[13]	AHP	AHP	Yes	No	No
[14]	User Level, Task urgency, Task Load and Time Queuing up	No	Yes	No	No

IV. PROPOSED PRIORITY TASK SCHEDULING STRATEGY

A. Priority Task Scheduling

In the literature and according to [22], a task is a single process or multiple processes which will be executed on a compute node presented by a VM in the cloud computing. However, Scheduling is a group of mechanisms that manage the order of execution of multiple tasks on the resources [5].

Task scheduling is divided into two categories: static scheduling and dynamic scheduling. In the static scheduling, the scheduling decisions are taken before tasks are submitted. The dynamic one is divided into two types: batch mode or online mode. In batch mode tasks are queued, and scheduled after fixed period. In online mode task is scheduled dynamically when they arrive in the system [23].

Our priority task scheduling strategy can be considered as dynamic as well as centralized scheduling.

The following figure presents the components of the proposed cloud controller, which is the most important component in this strategy.

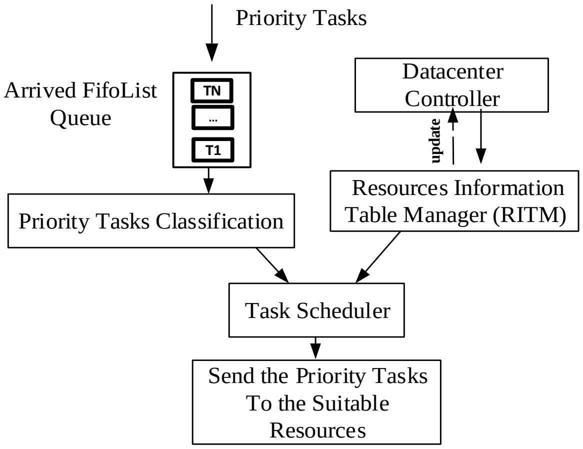

Priority Tasks
Arrived FifoList Queue
TN
...
T1
Datacenter Controller
update
Resources Information Table Manager (RITM)
Priority Tasks Classification
Task Scheduler
Send the Priority Tasks To the Suitable Resources

Fig. 2. Cloud Controller Process

As shown in the figure two, the priority task scheduling in the cloud controller is based on the four following steps.

1) Classification of VMs
2) Priority Tasks Classification
3) Assignment of the tasks to the VMs (Task scheduler)
4) Updating the resources information table

Based on those steps, the cloud controller assign tasks to the appropriate resources.

Here some abbreviations used in this paper:

TABLE II. ABBREVIATIONS TABLE

Notation	Definitions
Ti[Tl]	Length of a task i or the number of instruction that need the task to be executed in a resource.
Ti[Td]	The deadline of the tasks (given from the user SLA).
Ti[Ta]	The age of the task (the task waiting in the system).
QP	Priority Queue
SDC	The data-center speed.
SCL	The cluster speed.
SN	The node speed
P	The processing speed or ability in MIPS For The VM.
Load	The resource load

B. Priority Tasks Classification

In the cloud computing, different CSC generates different tasks having different length, different deadline, different arrived time and the sequence is dependent on the arriving time. In order to have an efficient Priority task scheduling strategy the priority tasks classification is very important. Various parameters can be used like:

- CPU Task utilization.
- RAM Task utilization.
- Bandwidth Task utilization.
- Task length or size.
- File Task size.
- The task deadline.
- The task waiting time in the system (Age).
- …..etc.

In our strategy, we are oriented to use dynamic priority based on three parameters: the task deadline, the task length, and the task age.

1) Task deadline: We have chosen the deadline as priority parameter because it is one of the important parameter that the CSP have to respect in order to respect the QoS that are documented in the SLA.

2) Task age: We have chosen the age as priority parameter because when the task has low priority, it has to wait for a long time and this leads to increase in execution time.

3) Task length: We have chosen the length as priority parameter because when the task has small lengths it can execute rapidly, thus, liberate the resources as soon as possible for the other tasks.

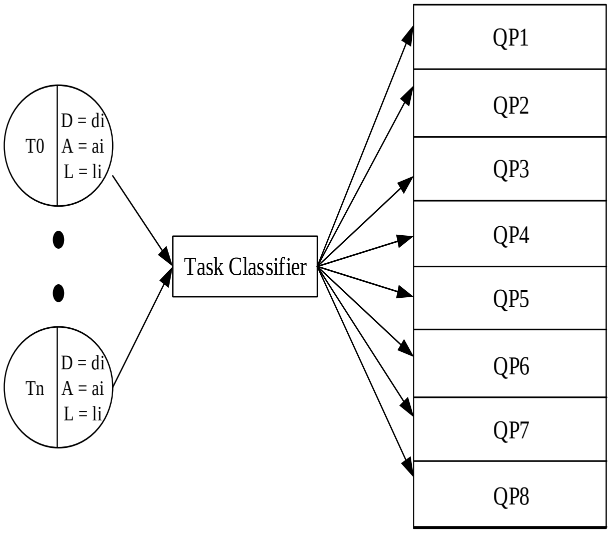


Fig. 3. Tasks Classification

As shown in the figure three, the task classifiers receive the priority tasks in order to classify them in the priority queues (from QP1 to QP8) and will work as follow:

1) Tasks having shorter deadline, shorter length and highest age will reside at the first priority queue (QP1).

2) Tasks having shorter deadline, shorter length and shorter age will reside at the second priority queue (QP2).

3) Tasks having shorter deadline, highest length and highest age will reside at the third priority queue (QP3).

4) Tasks having shorter deadline, highest length and shorter age will reside at the fourth priority queue (QP4).

5) Tasks having highest deadline, shorter length and highest age will reside at the fifth priority queue (QP5).

6) Tasks having highest deadline, shorter length and shorter age will reside at the sixth priority queue (QP6).

7) Tasks having highest deadline, highest length and highest age will reside at the seventh priority queue (QP7).

8) Tasks having highest deadline, highest length and shorter age will be present at the last priority queue (QP8).

The algorithm pseudo-code used in this step is explained in detail below.

ALGORITHM 1: Priority Tasks Classification

```
Repeat
For i from 1 to Arrive FifoList Queue Size do
Switch (Ti [Td] && Ti [Tl] && Ti [Ta]) {
Case Ti [Td] = Min_Deadline && Ti [Tl] = Min_Lenght && Ti [Ta] = Max_Age: Then put Ti in QP1
Case Ti [Td] = Min_Deadline && Ti [Tl] = Min_Lenght && Ti [Ta] = Min_Age: Then put Ti in QP2
Case Ti [Td] = Min_Deadline && Ti [Tl] = Max_Lenght && Ti [Ta] = Max_Age: Then put Ti in QP3
Case Ti [Td] = Min_Deadline && Ti [Tl] = Max_Lenght && Ti [Ta] = Min_Age: Then put Ti in QP4
Case Ti [Td] = Max_Deadline && Ti [Tl] = Min_Lenght && Ti [Ta] = Max_Age: Then put Ti in QP5
Case Ti [Td] = Max_Deadline && Ti [Tl] = Min_Lenght && Ti [Ta] = Min_Age: Then put Ti in QP6
Case Ti [Td] = Max_Deadline && Ti [Tl] = Max_Lenght && Ti [Ta] = Max_Age: Then put Ti in QP7
Case Ti [Td] = Max_Deadline && Ti [Tl] = Max_Lenght && Ti [Ta] = Min_Age: Then put Ti in QP8}
Remove Ti from FIFOList
End for
Until FifoList Queue is empty
```

C. VMs classification

As shown in the figure two, RITM obtains information about all resources through calling the datacenters controller, and each datacenter controller has a table that contains all information about their clusters (through calling their cluster controller) and so on until having information about all VMs. Hence, when the RITM receives the tables from the datacenters controller, it makes an update in its general table. This table contains for each resource its load and speed (performance). At last, it provides to the task scheduler a SVMT (Sorted VM Table) containing all VMs sorted according to their load.

TABLE III. SORTED VM TABLE (SVMT)

VMs (Sorted According To Their Load)
VMxxxx
…
VMxxxx

Note that the first x mean in which data-center the VM is created, the second mean in which cluster, the third in which node and finally the last x mean the number of this VM.

Having this information the task scheduler can choose the VM with the best features to execute a priority task. Moreover, it can choose the location of the data-center if required in the SLA contract.

D. Tasks Assignment to VMs

The pseudo-code of the priority task-scheduling algorithm is as a follow:

ALGORITHM2: Update Task Priority

```
Repeat
For i from 2 to 8 do
    For j from 1 to QPi.Size do
        If (Tj[Td] = Min_ DeadlineQPj) do
            Put Tj in the end of PQi-1
    End for
End for
```

```
8. Until all priority queues are empty
```

The main priority task scheduling algorithm is as follows:

ALGORITHM 3: Priority Task Scheduling

```
1. Repeat
2.   For j from 1 to 8 do
3.     If (QPj is not empty) do
4.       i=j
5.       Repeat
6.         Assign the first task in the QPi to the first VM in SVMT
7.       Until QPi is empty
8.     End if
9.   End for
10.  Update Task Priority
11.  Update SVMT
10. Until all priority queues are empty
```

Note that our algorithm offers several tasks priorities and this allows scheduling the highest priority tasks first. In the other side for the tasks that have least priority haven't waited a lot in their queues because they can ascend to the above queue thus have more priority.

E. Updating the resources information table

As shown in the figure two, the RITM is the only component that provides the resources information to the task scheduler. So, as to assign tasks to the resources (VMs). The RITM update dynamically its resources information table after a periodical time which can be defined either by the CSP or if there is a request from the task scheduler. Here, an example of this table.

TABLE IV. RESOURCES INFORMATION TABLE

Datacenters	**Clusters**	**Nodes**	**VMs**
Datacenter (SDCi,loadi)	$\text{Cluster}_{ij}\text{SCL}_{ij}$, Load_{ij})	Node_{ijk}(SN_{ijk}, load_{ijk})	VM_{ijkl}(P_{ijkl}, Load_{ijkl})

Note that the speed of each component is calculated through the speed addition of the under components and it is updated dynamically after a periodical time.

V. DISCUSSION AND CONCLUSION

Task scheduling is a big challenging issue in the cloud computing. In order to satisfy both the CSC as well as the CSP, efficient task scheduling strategy is required. The rest of this section is devoted to make a conclusion and a discussion about the proposed priority task scheduling strategy. Having examined existing scheduling algorithms that centered on priority, allow us to make a conclusion that mostly authors do not give an overview of task scheduling from the first step, when the task arrive, to the last step when the tasks are executed in the resources. They didn't give the importance to the priority tasks, in spite of, having a big importance in the SLA contract. In this paper, we have considered that each CSP has a multi-cloud data-center, each data-center has several clusters. As centralized and dynamic scheduling, the cloud controller is the only component who decides in which resources the priority tasks will be executed. It allows having a global view of all cloud resources components, which can even make a solution for the load balancing problems. Priority queues have also been considered for the tasks classification (from QP1 to QP8), according to three parameters: task deadline, task length and task age. Moreover, the task ascending in the priority queue also has taking into consideration, such as a task can ascend to the above queue if the task deadline equal to the min-deadline compared to the existing tasks deadline in the same queue at a specific time. Those features are the strength of our strategy that accelerated the execution of the priority task compared with the other one. Finally, as future work, our research continues by making the simulation for proposed priority task scheduling strategy, make in consideration the load balancing, and migration of running tasks from one machine to other machines with better performance.

REFERENCES

[1] Michael Armbrust, Armando Fox, Rean Griffith,Anthony D. Joseph, Randy Katz, Andy Konwinski,Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia "A View Of Cloud Computing" Communications Of The ACM – Vol. 53 (4), April 2010.

[2] Amardeep Singh And Inderjit Kaur "A Survey On Cloud Computing And Various Scheduling Algorithms" International Journal Of Advance Research In Computer Science And Management Studies (IJARCSMS) – Vol. 4 (2), February 2016.

[3] Definition of Cloud Computing, https://www.nist.gov/programs-projects/cloud-computing, viewed 07/12/2016.

[4] Raja Manish Singh, Sanchita Paul and Abhishek Kumar "Task Scheduling In Cloud Computing : Review",International Journal Of Computer Science And Information Technologies (IJCSIT) – Vol. 5 (6) , 2014.

[5] Amanpreet Kaur And Usvir Kaur "A Survey For Task Scheduling In Cloud Computing" International Journal Of Advanced Research In Computer Science And Software Engineering (IJARCSSE) – Vol. 6 (5), May 2016.

[6] Tingting Wang, Zhaobinliu , Yi Chen, Yujie Xu and Xiaoming Dai "Load Balancing Task Scheduling Based On Genetic Algorithm In Cloud Computing" The 12th IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC), 2014.

[7] Atul Vikas Lakraa and Dharmendra Kumar Yadavb "Multi-Objective Tasks Scheduling Algorithm For Cloud Computing Throughput Optimization" International Conference on Intelligent Computing, Communication & Convergence (ICCC), 2015.

[8] Samia Ijaz, Ehsan Ullah Munir, Waqas Anwar And Wasif Nasir "Efficient Scheduling Strategy For Task Graphs In Heterogeneous Computing Environment " The International Arab Journal of Information Technology (IAJIT), Vol. 10 (5), September 2013.

[9] Amit Agarwal And Saloni Jain "Efficient Optimal Algorithm Of Task Scheduling In Cloud Computing Environment" International Journal of Computer Trends and Technology (IJCTT) – Vol. 9 (7), March 2014.

[10] M. Lawanya Shri, M.B.Benjula Anbumalar, K. Santhi And Deepa.M "Task Scheduling Based On Efficient Optimal Algorithm In Cloud Computing Environment" International Conference on "Recent Research Development in Science, Engineering and Management (ICRRDSEM), May 2016.

[11] Deepika Saxena, R.K. Chauhan And Ramesh Kait "Dynamic Fair Priority Optimization Task Scheduling Algorithm In Cloud Computing: Concepts And Implementations", I. J. Computer Network and Information Security (IJCNIS), February 2016.

[12] Aditi Sharma And Shivi Sharma "Credit Based Scheduling Using Deadline In Cloud Computing Environment", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) – Vol. 4 (2), February 2016.

[13] Shamsollah Ghanbari And Mohamed Othman "A Priority Based Job Scheduling Algorithm In Cloud Computing" International Conference on Advances Science and Contemporary Engineering (ICASCE), January 2012.

[14] Pankajdeep Kaur And Parampreet Singh "Priority Based Scheduling Algorithm With Fast Task Completion Rate In Cloud" Advances In Computer Science And Information Technology (ACSIT) – Vol. 2 (10), April-June 2015.

[15] Er-raji Naoufal, Faouzia Benabbou And Ahmed Eddaoui "Task Scheduling Algorithms In The Cloud Computing Environment: Survey And Solutions" International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) – Vol. 6 (1), January 2016.

[16] S. Jeyalakshmi And N. Sankarram "Scheduling Algorithms For Cloud Computing Environments And Research Issues" Journal Of Applied Sciences Research (JASR) – Vol. 12 (3), March 2016.

[17] Jyoti Thaman And Manpreet Singh "Current Perspective In Task Scheduling Techniques In Cloud Computing: A Review" International Journal in Foundations of Computer Science & Technology (IJFCST) – Vol. 6 (1), January 2016.

[18] Rupinderjit Singh And Er.Manoj Agnihotri "A Review Of Cloud Computing Scheduling Strategies" International Journal Of Engineering Trends And Applications (IJETA) – Vol. 3 (4), July-August 2016.

[19] Amardeep Singh And Inderjit Kaur "A Survey On Cloud Computing And Various Scheduling Algorithms" International Journal Of Advance Research In Computer Science And Management Studies (IJARCSMS) – Vol. 4 (2), February 2016.

[20] Sujit Tilak And Prof. Dipti Patil "A Survey Of Various Scheduling Algorithms In Cloud Environment" International Journal Of Engineering Inventions (IJEI) – Vol. 1 (2), September 2012.

[21] Swachil Patel And Upendra Bhoi "Priority Based Job Scheduling Techniques In Cloud Computing: A Systematic Review" International Journal Of Scientific & Technology Research (IJSTR) – Vol 2 (11), November 2013.

[22] Definition Of Tasks, https://msdn.microsoft.com/en-us/library/bb525214(v=vs.85).aspx, viewed 07/12/2016.

[23] P. Akilandeswari and H. Srimathi "Survey And Analysis On Task Scheduling In Cloud Environment" Indian Journal of Science and Technology(INDJST) – Vol 9 (37), October 2016.

Sentiment Analysis Challenges of Informal Arabic Language

Salihah AlOtaibi
Information Systems Department,
College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, KSA

Muhammad Badruddin Khan
Information Systems Department,
College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, KSA

***Abstract*—Recently, there are wide numbers of users that use the social network like Twitter, Facebook, MySpace to share various kinds of resources, express their opinions, thoughts, messages in real time. Thus, increase the amount of electronic content that generated by users. Sentiment analysis becomes a very interesting topic in research community. Thereby, we need to give more attention to Arabic sentiment analysis. This paper discusses the challenges and obstacles when analyze the sentiment analysis of informal Arabic, the social media. The most of recent research sentiment analysis conduct for English text. Also, when the research works in Arabic sentiment analysis, they focus in formal Arabic. However, most of social media network use the informal Arabic (colloquial) such as Twitter and YouTube website. This paper investigates the problems and the challenges to identify sentiment in informal Arabic language which is mostly used when users express their opinions and feelings in context of twitter and YouTube Arabic content.**

Keywords—Informal Arabic; Sentiment analysis; Opinion Mining (OM); Twitter; YouTube

I. Introduction

Arabic is a Semitic language spoken by more than 330 million people as a native language. Arabic is a highly structured and derivational language, in which morphology has a very important role. Thus, Arabic natural language processing (NLP) applications must deal with the complex nature of the Arabic language. For example, Arabic is written from right to left and no capitalization is used for nouns, which is a necessary feature in text mining. The Arabic language contains 28 letters and, in addition, the Hamza (ء). In Arabic, letters change their shape according to their position in the word (beginning, middle, or end) [1]. For example, see letter " ي/ya'a" and letter "ج/geem," as shown in Table I.

TABLE I. Position of the Character in the Word

Letter	Beginning	Middle	End
ي	يـ	ـيـ	ـي
ج	جـ	ـجـ	ـج

Arabic is the official language of Islam and of the last Prophet. It was selected to be the language of the Holy Qur'an. Muslims living throughout the world, thus, feel an affiliation with the Arabic language[1].

A. *Types of arabic*

There are three types of Arabic: Classical Arabic language (CA), Modern Standard Arabic language (MSA), and informal Arabic language (the latter is sometimes referred to as colloquial Arabic language).

CA is the language of Islam, which Arabic speakers use in their prayers and when reading the Qur'an. MSA is the official language across the Arab world. It is used by educated people in more formal circumstances; for example, for news reports, in classrooms, and in business. Informal Arabic is the language that people speak daily with family and friends, in which people also use their own dialects, which vary from region to region. The three different styles of Arabic language are available to every Arab – for example, each day, an Arabic speaker will use Classical Arabic for his five daily prayers, MSA when listening to or reading the news, and his/her own dialect when at home. Each type of Arabic has its own grammar, lexicon, and morphology, although though some properties are shared between the varieties. Most existing research tools have been developed to handle text that is written in MSA. This constitutes a limitation when it comes to research that focuses on text mining in relation to informal Arabic language [1], [2], [3].

In the research field, the sentiment analysis becomes hot topic to work in. The most of research and techniques for sentiment analysis is for English text. Thereby, it is obvious, there are limitations in the researches that interest for sentiment analysis for Arabic language [4]. Moreover, most of the researchers focus on formal Arabic language [5]. Since most of users use informal Arabic in the world of social media, the task of sentiment analysis becomes more sophisticated [6]. This motivates us to explore the challenges to analyze the sentiment for informal Arabic language such the different Arabic Dialects are another challenge.

The paper is organized in few sections to describe further details of our work. Section 2 describes the nature and the complexity of Arabic language. Section 3 gives overview about the main commons and differences between informal Arabic and informal English. Section 4 gives overview about the challenges in sentiment analysis for Arabic language. In section 5, outlines the related work done in this area. In section 6, gives overview about the main commons and differences between Twitter and YouTube dataset. In section 7, we describe the method and the preprocessing. Section 8 shows our finding and discussion. Finally, in the brief Section 9, we make concluding remarks.

II. The Complexity of Arabic Language

The Arabic language is challenging and complex due to its nature and characteristics. The following paragraphs illustrate the complexity of Arabic.

This section provides a literature review for the field of sentiment and semantic analysis, focusing mainly on informal Arabic language.

A. Word meaning

The term "word" defines a single, isolated item between two spaces, which has a certain meaning. In Arabic, it is common for one word to have several different meanings, depending on the context. Table II gives the example of the Arabic word سهل/sahel, which can be used as a noun with three different meanings. The phrases have been taken from Twitter [7].

TABLE II. Meanings of the Word سهل/Sahel as a Noun

Sentences	Phrases in English	Word Meaning
سهل منبسط *sahel moonbaset*	"flat plain"	Flat floor
سهل بن سعد *Sahel bin Saad*	"Sahel bin Saad"	Name
سَهْلُ المَرام *sahel almaraam*	"easy to get"	Easy

B. Variations in lexical category

In Arabic linguistics, a word can be a noun, verb, or particle. The term "particle" covers all other words that are not nouns or verbs, such as prepositions and conjunctions, for instance. Examples are given in Table III.

TABLE III. Word Types in the Arabic Language

Word Type	Example	English Translation
Noun	كتاب	Book
Verb	يكتب	Write
Particle	على	On

Moreover, a word can belong to different lexical categories, depending on the context. Table IV shows how the word حلق/halq can be used in different parts of speech [7].

TABLE IV. Lexical Categories for the Word حلق/Halq

Phrases	Phrases in English	Word Category	Word Meaning
حلّق الانسان *halq alensan*	"human throat"	Noun	Throat
حَلَق رأسه *halq ra'asah*	"shaving his head"	Verb	Shaving
حَلّق الطائر *halq alta'er*	"flying bird"	Verb	Fly

C. Morphological characteristics

Morphology is a branch of linguistics that deals with the structure of words. It concerns word formation, roots, and affixation behaviors. Arabic is a highly structured and derivational language. Arabic is a Semitic language and it is morphologically complex. Typically, a word in a Semitic language contains more information than a word in a non-Semitic language like English.

In Arabic, for example, various affixes can be attached to create new words; from the root word درس/*darasa*, for instance, several different words can be generated, such as يدرس/*yadras* ("studying" in English), مدرس/*modras* (English: "teacher"), مدرسة/madrasa (English: "school"), and مدارس/*madares* (English: "schools") [8]. Below is short description of each basic item in the Arabic language.

As clarified above, a word is a single, isolated item with a certain meaning. In Arabic, a word can be a noun, verb, or particle, and the same word can fit into different categories, depending on the context.

A morpheme is the smallest linguistic unit that has a meaning. A morpheme cannot be split into smaller units. Morphemes should give a meaning to the word of which they are a part.

A root is a single morpheme that provides the basic meaning of a word. In Arabic, the root is the original form of the word, before any transformation process occurs. Many words can be formed using one root.

A *stem* is a morpheme without an affix. The stem provides a specific idea or meaning. In English, the root is also sometimes called the "stem" or "word base," but in Arabic, the stem (or base) is different from the root [7]. Table V illustrates the morphological characteristics of Arabic.

TABLE V. Morphological Characteristics

Morphological characteristics	Definition	Example
Word	a single and isolated item between two spaces	المحمدون/alMuhammadwn
Morpheme	smallest linguistic unit that has a meaning	ـون / Wn
Stem	The basic form of word	محمد /Muhammad
Root	The original form of word	حمد/Hammd

An *affix* is a morpheme that can be added before (*prefix*), after (*suffix*), or within (*infix*) the root or stem to give a new word or meaning [7]. Table VI shows how the word سجد/ Sajed can have different meanings when various affixes are added.

TABLE VI. Different Meanings of سجد/ Sayed When Different Affixes Added

Word	English translation	Suffix	Infix	Prefix
ساجد/Sajeed	"Prostrate"	***	ا	***
مسجد/ Msadjad	"Msadjad"	***	***	م
سجادة/Sejada	"Carpet"	دة	ا	***

In text mining, the stemming process is usually used to convert a word into its root form. The main objective of the stemming process is to remove all possible affixes, thus diminishing the complexity of a word and reducing the number of features and tokens in corpora [7]. For example, if the words ذاهبون/*thahebon*, ذهبوا/*thahabo*, and يذهب/*yathhab* are all in a corpus, after the stemming process has taken place, all the words will be recognized in the text mining procedure as the same word ذهب/*thhab* ("go" in English). However, the stemming process is not always considered beneficial in

Arabic because the Arabic root is context dependent; thus, a stem may lead to more than one definition [9]. Table VII exemplifies words with different meanings that share a common root.

TABLE VII. DIFFERENT WORDS WITH THE SAME ROOT

Sentences	English translate	Root	Meaning
يخرج من المنزل *yakrooj men almanzel*	“leaves home”	خرج	Goes out
تخرج من الجامعه *takarraj men aljameea*	“graduates from college”	خرج	Graduate

Vowelization or *diacritization* is the process of putting diacritical mark vowels above or under letters in Arabic words (*fatha*: ◌َ, *dammah*: ◌ُ, *kasrah*: ◌ِ). *Nunation* is the process of putting a set of diacritically marked vowels at the end of a word to create the sound of the letter ن/*N*. The *kasheeda* (—) or *tatweel* is the symbol used to stretch some Arabic characters [7]. The *tatweel* symbol is often used in informal Arabic language to emphasize a feeling or meaning. In the text mining process, the *tatweel* must be removed because it creates multiple forms of the same word. Table VIII shows how *tatweel* preformed different forms for one words.

TABLE VIII. TATWEEL

Word	English translation
مرحبا مرحبـا مرحبــــا مرحبـــــــــا / marhaba	“hello”

III. THE INFORMAL ARABIC VS. INFORMAL ENGLISH LANGUAGE

Informal language could be described as language that ignores the standard rules of grammar and spelling. In general, the Arabic language is written from right to left, while English is written from left to right. There is no capitalization in Arabic, unlike in English [1].

Informal English uses abbreviations (for example, “m8” for “mate” and “u” for “you”), whereas in Arabic, there are no such abbreviations. In informal Arabic language, abbreviations called *Arabization* are used (like برب for “be right back” and لول for “laughing out loud”). Arabization is the process of translating new concepts and terminology into Arabic. In fact, with Arabization, users translate only the first letter of each word in the English phrase or sentence to create a new abbreviation in Arabic (so, using the previous example of “be right back,” برب is “BRB”). The main commonalities between informal Arabic and informal English are the use of emoticons, texting-style abbreviations, and repeated letters or punctuation, which is added for emphasis [10].

IV. ARABIC SENTIMENT ANALYSIS CHALLENGES

NLP for Arabic is fraught with many challenges, some of which result from the structural and morphological complexity of the language. As mentioned previously, Arabic is a derivational language, which means that many words can be formed from three-letter roots. The resulting words may look similar, but have very different meanings. Arabic grammar is also highly complex, containing a variety of sentence structures, both verbal and nominal. A verbal sentence is one that starts with a verb phrase, whereas a nominal sentence starts with a noun phrase. Arabic also contains many word forms and diacritics [1], [4]. The complex features of the language make the task of analysis more difficult [11]. Furthermore, the semantic dictionaries or lexicons on offer for Arabic text analysis are limited. Indeed, future research should consider the necessity of creating morphological analysis tools for Arabic text analysis that can cover all word forms and can perform suffix, affix, prefix, and root extraction. Grammatical analyzers and/or part-of-speech (POS) taggers are also needed. Some morphological analyzers have been developed for use with the Arabic language, such as BAMA (the Buckwalter Arabic Morphological Analyzer) and MADA (the Morphological Analysis and Disambiguation for Arabic analyzer). There are no sophisticated POS taggers and lexicons tools in Arabic which identify all parts of speech and discover the difference of sentence's types. These issues present a challenge for sentiment mining, which generally requires both semantic analysis of words and grammatical analysis of text [4].

In fact, another major challenge that has surfaced due to the emergence of social media is that most of the Arabic language found on the internet is written in informal Arabic. The informal version of the language is unstructured in nature. Furthermore, many users utilize their own regional dialects, rather than opting for modern standard Arabic; for instance, the word شوف/*shoof*, which means “look” in English, might be used instead of the word أنظر/*onther*. Another important point is that informal Arabic does not use diacritics; thus, in some cases, the meaning of the word becomes ambiguous. For example, the words مُدَرِّسة (“teacher”) and مَدْرسَة (“school”) look the same when written without diacritics (مدرسة). Social media has also given rise to the increased usage of letter repetition to emphasize the meaning or feeling associated with a word (شكراااااا – “thankssss,” as opposed to شكرا – “thanks”) [12].

Informal Arabic words usually do not have their own specific roots. Indeed, a stemmer will sometimes identify the same root for both the informal word and the formal word, as is the case with the terms راحه/*rahaah* (formal) (“comfort” in English), andنروح /*nrooha* (informal) (“go” in English), both of which take the root روح/*rooh* [13]. Another key trait in Arabic social media is the use of compound phrases and idioms to express opinions; e.g., يا ولد يا مطوع/*ya walad ya motaua* (a negative expression that belittles someone pretending to be religious). Compound phrases and idioms vary from one country to another. Also, that gives different sentiment polarities rather than its constituent words itself. According to previous examples, the sentiment polarity is negative while none of its constituent words are negative [14].

As most social media users utilize informal Arabic, the task of text analysis therefore becomes more challenging. The introduction of various dialects poses a further difficulty [6] as does the lack of literature on informal Arabic language [5]. These factors motivated us to focus on the problems that exist in informal Arabic, with the aim of encouraging more researchers to participate in this field.

V. Related Work

Sentiment analysis depends on using various techniques of machine learning, such as Knowledge-based, corpus-based, Naïve Bayes (NB), support vector machine (SVM) and maximum Entropy model (ME). Sentiment analysis can be applied on different types of content such as content of newspapers, review sites, tweets from twitter site [15].

A. *Sentiment analysis on Arabic Content*

The sentiment analysis for Arabic language became topic of interest for many researches to participate in this field. In one study, researchers presented an advanced technique for inferring sentiment orientation of social media sites focusing on the problems related to web dependent analysis [16]. New tool was developed that can be used for Arabic sentiment analysis. The proposed tool is divided into two techniques; NLP and human computation. The proposed system consists of two parts; game-based lexicon and sentiment analyzer parts. The first part is used to build the lexicon based on human computation, while the second part is a sentiment analyzer that takes each review and executes sentences segmentation [5].

Other researchers proposed a new technique for Sentiment Analysis and Subjectivity Analysis (SSA) for certain Arabic social media sites. Results demonstrated that the use of lexeme or lemma data is useful. On the other hand, there is a need for individualized solutions for every task and genre [8]. Also, there is research work performed to do the sentiment analysis for Arabic Facebook news pages. They used three machine learning classification techniques; Naive Bayes, SVM and decision tree are used to improve the sentiment analyzer [17]. Some researchers also, proposed a technique for extracting and analyzing Arabic business reviews that are available in forums and blogs. The system has two basic parts; reviews classifier and sentiment analyzer. First part classifies the web page. Second part for detecting the polarity of the sentences based on an Arabic lexicon [18]. In 2012 an advanced Arabic sentence level sentiment categorization technique was introduced that depends on two methods; a grammatical and semantic methods. [19].

B. *Arabic Sentiment analysis on twitter*

As we mentioned in previous paragraphs, the research on Arabic semantic is limited. One of those limited studies was provided by A. Shoukry and A. Rafea. They produce an application on Arabic sentiment analysis by classification the Arabic tweets. They used different ML classifiers and different features. They apply the SVM and naïve bayes and also try the combinations of classifiers [3]. Also, other researchers tried to find and explore the problems of sentiment analysis for informal Arabic. They apply their experiments on twitter. They use knowledge-based technique. There is a limitation in the number of Arabic sentiment lexicons, and the main challenge is to build lexicons for informal words [13].

VI. Twitter Data Vs. YouTube Data

Twitter is a microblog and social network that allows users to share their thoughts and express their opinions through short massages. While YouTube is a website designed for sharing video. In YouTube the users can restrict who views their videos with YouTube's privacy option. Also the users can post a comment and reviews on the videos that were viewing. There is some common and different between Twitter and YouTube Arabic text.

The most commons between Twitter and YouTube users' post are all of the users use informal language that ignores the standard rules of grammar and spelling. Also the posts contain emoticons, texting-style abbreviations, and repeated letters or punctuation added for emphasis.

The main differences, on Twitter, users produce short pieces of information known as "tweets" (limited to 140 characters). One can find a diverse range of topics within these tweets. Twitter users may post tweets expressing opinions about personalities, politicians, products, companies, and events, for instance [20], [21], [22]. Furthermore, some of the symbols used in tweets are language-independent. For example, "@" is utilized when users are referring to other users. "#" (hash tag) is used to mark topics or keywords—it is used to make messages more visible to other people. "RT" (re-tweet) is used when someone likes a tweet and wants to repeat it for their followers to see. The writing technique for tweets is fast and short. Users utilize acronyms and emoticons to express their opinions.

On YouTube, users produce reviews and opinions on contains of videos. There is no limited length for reviews posts. The posts only reviews or comment on contains of videos unlike the twitter tweets. There are no special symbols used in reviews like tweets.

VII. Method

This paper aims to investigate the problem and challenges of informal Arabic sentiment analysis. In this paper, we used twitter and YouTube datasets. The processing of the method can be described as follows: 1) after collecting the datasets, we determine the annotation of each tweets and each YouTube review (positive, negative, and neutral). 2) Convert the emotion icons to text. 3) Clean the dataset by removing: names, URL, pictures, English word, for tweets re-tweets sign, hash tags. 4) Normalizing process which makes the text in consistent form, in other words, convert all different forms of word to a common form. 5) Tokenization process applied on each tweets to divide them into multiple tokens based on whitespaces characters. 6) Then make stemming process to return each word to its root. 7) Remove the Arabic stop-word. The result of preprocess is used as input to the classifier model to test the result. The sentiment classifier used in the model is Naïve Bayes algorithm.

VIII. Findings and Discussion

Informal Arabic language, in general, is "noisy" and poorly structured. It also features the non-standard repetition of letters, abbreviations, and emoticons, as well as the use of Arabized words.

Arabic tweets and YouTube reviews contain incorrect and misspelled word(s). These spelling problems needs special attention and require proper cleaning. When applying sentiment analysis for informal Arabic many problems occurred in text processing step. There are various problems that were found in each text processing phase. The following sub-sections expound the problems in each phase:

A. *Tokenization phase*

When applying sentiment analysis for informal Arabic many problems were encountered. The problems explained as following

1) *Repetition Letters*

The first problem is the repetition of letters, as mentioned in section 4. As we know that in the Arabic language if we have repeated letters in the text it cannot occur more than twice. So if the repetition exists at beginning, middle or at the end of the word more than two times, it will be detected in the pre-processing step. Unfortunately, repetition cannot be detected where a letter is repeated only twice. Table IX shows pre-processing of tweets with repetition letters. In literature issue of detection of the repetition is discussed for situation with repetition only existing at the end of word [13].

TABLE IX. REPETITION LETTER PROBLEM

Platform	Sentences	English Translate	After pre-Processing
Twitter	كئيب ل ابعععععد حد Kaeeb le abeeed haad	I am very depressed	كئيب ل ابععد حد
Twitter	هههههههه جميل جداً :) hahahahah Jamel jeedan :)	Very beautiful :)	هه جميل جدا :)
YouTube	أحسسسسسسسسسسسسسسسن Ahsssssssan	Better	أحسسن
YouTube	بالصراحة مرة واوووووو beSaraha Marra wowwwww	In fact it is wow	بالصراحة مرة واوو

2) *Negations*

The second problem is that word polarities are affected significantly by ignoring negations like ما/*Ma*, لا/*Laa*, لم/*lam*, and لن/*lan* which are formal Arabic negations. The informal Arabic contains many of informal negation words like مو/*Muo*, مش/*Mush*, and موب/*Moub*, which also affect the text polarities by converting the meaning of the sentence to exactly the opposite. Furthermore, as we mentioned in section 3, the informal Arabic used Arabized words. The Arabized words "نو" and "نوت" which means in English "no" and "not", are also used as negations words in informal Arabic. Table X shows how the informal negation words affected the text polarities.

A negation indicator should, therefore, be used to detect polarities accurately.

TABLE X. WHO NEGATIONS AFFECT THE TEXT POLARITIES

Platform	Sentences	English Translate	Polarities	Sentences Without Negation	Polarities
Twitter	**مش** أهبل :) Mush Ahbal	Not idiot	positive	أهبل :)	negative
Twitter	ليش انا**مو** جريئه :(Leash Ina mu jareeah :(	Why I am not bold	negative	ليش انا جريئه :(	positive
Twitter	***نوو*** يفهم شي No yefham shee	Does not understand something	negative	يفهم شي	positive
YouTube	**مو** ملاهي ذيي Mu malahe thee	This is not amusement park	negative	ملاهي ذيي	positive
YouTube	انا**موب** طفل Ana moub teffel	I am not a child,	Positive	انا طفل انا 14 سنه	negative

3) *Connecting different words together*

The third problem involves Twitter users connecting different words together—this method of writing occurs frequently in tweets because the length of a tweet is limited. This issue affects stop-word filtering because certain stop words are not removed and new forms of words are created. Table XI illustrate how this problem affects the pre-processing step by increasing the number of tokens

TABLE XI. THE EFFECT OF CONNECTING DIFFERENT WORDS TOGETHER AT TOKENIZATION AND STOP WORDS FALTERING

Tweet Problem	Sentences	English Translate	Tokenizing Process	Falter Stop Word
Tweet contain connecting words together	يامرحبا تسلم عزيز ابوخالد Yaa marhaba teslam aziz ibu Kaled	Hello, thank you Ibu Kaled, you are dear and precious person	يامرحبا تسلم عزيز وغالي ابوخالد	يامرحبا تسلم عزيز وغالي ابوخالد
Tweet does not contain connecting words together	يا مرحبا تسلم عزيز ابو خالد Yaa marhaba teslam aziz wa ghali ibu Kaled		يا مرحبا تسلم عزيز ابو خالد	مرحبا تسلم عزيز خالد

From the table above, shows the results of tokenization process and faltering the stop words are different based on how the tweet is written.

Connecting different words together can also cause ambiguities in meaning like words وفي/*wafee* and وهم/*whum* have two different meanings with/without connection as can be seen in Table XII.

TABLE XII. THE CONNECTING DIFFERENT WORDS TOGETHER CAUSE AMBIGUITIES IN MEANING

Platform	Sentences	English Translate	Word	Meaning
Twitter	تبتسم :)) . وفي عينيك ألف دمعة Tebtassen wafee Eaneek alf damaah	Smiling :)). and Thousands of tears in your eyes	و+في	And in
Twitter	قلبي وفي . ماني مثل غيري Galbe wafee mane methel garre	I have the loyal heart .. I am not like the other	وفي	Loyal
Twitter	الناس يغلطون وهم اللي يزعلون 🔪☹ Nass Agtaiwn waahum elle yezalon	people make mistakes and also they Angry	و+هم	And they
Twitter	وهم و حيره Waham waa heera	Illusion and confusion	وهم	worry
YouTube	وهم مايعترفوا بفشلهم Waa hom Mayereefo be fashalho	They did not admit for they failing	و+هم	And they
YouTube	وفي مقاطع فيديو رعب Waa fee makateea video roob	And in video clips horror	و+في	And in

4) Diacritization problem

The tokenization is performed based on finding whitespaces characters. Some types of punctuations like diacritic are removed and then add single space, so the word broken to many tokens. The problem was variations of word forms and diacritic. Table XIII shows the diacritic problems.

TABLE XIII. THE DIACRITIC PROBLEMS DURING TOKENIZATION PROCESS

Platform	Tweet before Tokenization	Tweet after Tokenization
Twitter	إن الصَلاةَ كَانت عَلَى الْمؤْمِنينَ كِتاباً موقوتًا “ena alsalat kanat ala almoemenen ketaaban moqouta”	إن الص لاة ك انت ع ل ى ال مؤ م نين ك تابا موقوت
Twitter	راح انَتْحـَر ساعدوني “Raah Inteheer saodony”	راح ان ت حـ ر ساعدوني

The problem of the deletion of diacritics and certain word forms, like tatweel cases, was discussed in section 2. The problem was solved in this study during the suggestion pre-processing stage. Table XIV shows the normalization cases that were used in pre-processing.

TABLE XIV. NORMALIZATION CASES

Rule	Example
Tashkeel	المؤمنين<-الْمؤْمِنينَ
Tatweel	الله<-اللـــــه
Alef	آor أor إ->ا
Heh	ﻩor ة-> ه

5) Emoticons problem

Informal Arabic language text often uses emoticons, which cannot be interpreted by text-based models.

When the text was filtered to remove English words and special characters, all the emoticons were also removed. Thus, to preserve the emoticons, meaningful names were given to each symbol appearing in the corpus, which allowed the role of emoticons to be examined at sentiment analysis model. Table XV shows examples of the emotion icons conversion step.

TABLE XV. EXAMPLES OF THE CONVERTING EMOTION ICONS TO MEANINGFUL TEXT

Platform	Emotion icons	Sentences	English Translate	After converting the icons
Twitter	♥)':	مع السلامه :') ♥ “Maa alslamah ♥)':”	Goodbye)': ♥	مع السلامه رمزحزين رمزقلب
Twitter	O_o	متردد O_o “Motaraded O_o”	Hesitant O_o	متردد رمزمتفاجئ
YouTube	(:	حلوه وتضحك (:هههههههه..	Sweet and laugh (:	حلوه وتضحك هههههه.. رمزمبتسم
YouTube	._.	غبي ._.	Dumbass ._.	غبي رمزمتفاجئ

6) Writing style in informal Arabic text

Some writing styles used in informal Arabic text can affect text pre-processing results, such as when a word is written inside another word, or write the word in separate letters to emphasize the meaning or feeling, as shown in Table XVI.

TABLE XVI. EXAMPLES OF WRITTEN STYLES USED IN INFORMAL ARABIC LANGUAGE, AND TOKENIZATION PROCESSING RESULTS

Problem	English Transla-tion	Formal Sentence Style	Tokeniz-ation Processin g Results	Informal Sentence Style	Tokeniz -ation Processi ng Results
Writing a word inside another word.	Welcome	اهلا و سهلا *Ahlan wa sahlam*	اهلا و سهلا	اهـ وسهلا ـلا *Ahlan wa sahlam*	اهـ وسهلا ـــلا
	Hello or *Aslam alukom*	السلام عليكم *Aslam alukom*	السلام عليكم	السـ عليكم ـلام *Aslam alukom*	الســ عليكم ــلام
Word with separate letters	I finished	أنا منتهية *Ana Mentahea*	أنا منتهية	أنا م ن ه ي ه *Ana Mentahea*	أنا م ن ه ي ه

B. *Filter Arabic stop words phase:*

There is no given stop word list for informal Arabic language which contain informal Arabic words like: هاذي/*hathe*, هاذا/*hatha*,دي /*dee*, اللي/*elle*, so we build our own stop word list for informal Arabic language.

C. *Stemmer phase:*

In the Arabic there are different words with different meaning have the same root. This makes detecting the polarities of these words incorrect. As we mentioned above, in section 3.

Also other problem occurs during the stemming process. The stemmer some time deleted some basic letters the word Table XVII shows the light stemmer problems. We remove the stemmer step from the text processing.

TABLE XVII. STEMMER DELETED SOME BASIC LETTER FROM THE WORD

Platform	Sentences	English Translate	Stemmer results
Twitter	القران الكريم❤ *AL Quran al Kareem*	Koran Kareem ❤	القران → قر
Twitter	انزين انتهينا *Enzaeen entahena*	now we finished	انزين → انز
Twitter	يا الله ساعدني *Ya Allah saedney*	O God, help me	الله → له
YouTube	يا الله مقرف *Ya allah mogreef*	O God, disgusting	الله → له
YouTube	هذا فلم كرتون *Hatha film carton*	This film carton	كرتون→كرى

IX. CONCLUSION

The Arabic language is both challenging due to its complex linguistic structure and interesting because of its history and importance in religion, culture, and literature. Informal Arabic language, in general, is "noisy" and poorly structured. It also features the non-standard repetition of letters, abbreviations, and emoticons, as well as the use of Arabized words. Thus, these features should be considered during text mining. This paper investigates the problems and the challenges to identify sentiment in informal Arabic language in context of twitter and YouTube Arabic content. In this experiment, we found many issues that can be motivating for future research

REFERENCES

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 4, p. 14, 2009.

[2] M. Korayem, D. Crandall, and M. Abdul-Mageed, "Subjectivity and sentiment analysis of Arabic: A survey," in *Advanced machine learning technologies and applications*, vol. 322, Berlin & Heidelberg, Germany: Springer, 2012, pp. 128–139.

[3] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering," *Int. J. Data Min. Knowl. Manag. Process*, vol. 3, no. 1, pp. 79–95, 2013.

[4] N. Farra, E. Challita, R. A. Assi, and H. Hajj, "Sentence-level and document-level sentiment mining for Arabic texts," in *2010 IEEE international conference on data mining workshops (ICDMW), 13 Dec. 2010*, 2010, pp. 1114–1119.

[5] A. A. Al-Subaihin, H. S. Al-Khalifa, and A. S. Al-Salman, "A proposed sentiment analysis tool for modern Arabic using human-based computing," in *Proceedings of the 13th international conference on information integration and web-based applications and services*, 2011, pp. 543–546.

[6] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *2012 international conference on collaboration technologies and systems (CTS), 21-25 May 2012*, 2012, pp. 546–550.

[7] I. A. Al-Sughaiyer and I. A. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, no. 3, pp. 189–213, 2004.

[8] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 20–37, 2014.

[9] A. Moh'd Mesleh, "Support vector machines based Arabic language text classification system: Feature selection comparative study," in *Advances in computer and information sciences and engineering*, T. Sobh, Ed. Netherlands: Springer, 2008, pp. 11–16.

[10] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.

[11] P. Pak, Alexander and Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Lrec*, vol. 10, pp. pp. 1320–1326, 2010.

[12] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment Analysis in Arabic tweets," in *5th international conference on information and communication systems (ICICS), 1-3 April 2014*, 2014, pp. 1–6.

[13] L. Albraheem and H. S. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in *Proceedings of the 14th international conference on information integration and web-based applications and services*, 2012, pp. 415–418.

[14] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of arabic social media: A case study," in *2013 9th international conference on innovations in Information Technology (IIT), 17-19 March 2013*, 2013, pp. 215–220.

[15] A. Kumar and T. M. Sebastian, "Sentiment analysis on Twitter," *Int. J. Comput. Sci. Issues*, vol. 9, no. 4, pp. 372–378, 2012.

[16] R. Colbaugh and K. Glass, "Estimating sentiment orientation in social media for intelligence monitoring and analysis.," in *2010 IEEE international conference on intelligence and security informatics (ISI)*, 2010, pp. 135–137.

[17] A. E.-D. A. Hamouda and F. E. El-Taher, "Sentiment analyzer for Arabic Comments System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 3, pp. 99–103, 2013.

[18] M. Elhawary and M. Elfeky, "Mining Arabic business reviews," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 2010, pp. 1108–1113.

[19] M. Abdul-Mageed and M. T. Diab, "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.," in *LREC*, 2012, pp. 3907–3914.

[20] K. Makice, *Twitter API: Up and running: Learn how to build applications with the Twitter API*. O'Reilly Media, Inc., 2009.

[21] L. Barbosa and J. Feng, "Robust sentiment detection on Twitter from biased and noisy data," in *Proceedings of the 23rd international conference on computational linguistics: Posters*, 2010, pp. 36–44.

[22] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Proj. Report, Stanford*, pp. 1–6, 2009.

Need and Role of Scala Implementations in Bioinformatics

Abbas Rehman
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Muhammad Atif Sarwar
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Ali Abbas
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Javed Ferzund
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

***Abstract*—Next Generation Sequencing has resulted in the generation of large number of omics data at a faster speed that was not possible before. This data is only useful if it can be stored and analyzed at the same speed. Big Data platforms and tools like Apache Hadoop and Spark has solved this problem. However, most of the algorithms used in bioinformatics for Pairwise alignment, Multiple Alignment and Motif finding are not implemented for Hadoop or Spark. Scala is a powerful language supported by Spark. It provides, constructs like traits, closures, functions, pattern matching and extractors that make it suitable for Bioinformatics applications. This article explores the Bioinformatics areas where Scala can be used efficiently for data analysis. It also highlights the need for Scala implementation of algorithms used in Bioinformatics.**

***Keywords*—Scala; Big Data; Hadoop; Spark; Next Generation Sequencing; Genomics; RNA; DNA; Bioinformatics**

I. Introduction

Today, we are living in the world of Big Data. Huge amount of data is being produced on daily basis. Major sources of data include social media, enterprise systems, sensor based applications, Bioinformatics sequencing machines, smart phones, digital videos or pictures and World Wide Web. Big Data’s characteristics are Veracity, Velocity, Variety, Volume and Potential Value (these are known as 5 V’s). To make this data useful, it needs to be stored and analyzed with accuracy and speed. Traditional techniques are unable to store and analyze such large amount of data. These techniques are better for a limited amount of data analyses as the cost of analysis increases with increment in data volume.

To deal with this hurdle, Big Data platforms and tools are introduced which can analyze a large amount of data with accuracy, speed and scalability. Using Big Data Platforms like Hadoop, cost of analysis is also reduced as it runs on commodity hardware. Major challenges for Big Data are speed, performance, efficiency, scalability and accuracy. Big Data platforms and tools like Hadoop (distributed management System) and Apache Spark (for big data analysis) address these issues. NGS (Next Generation Sequencing) machines bring an evolutionary change in data generation of different sequences. NGS machines are generating a huge amount of sequence data per day that needs to be stored, analyzed and managed well to seek the maximum advantages from this. Existing bioinformatics techniques, tools or software are not keeping pace with the speed of data generation. Old Bioinformatics tools have very less performance, accuracy and scalability while analyzing large amount of data. When storing, managing and analyzing large amount of data which is being generated now a days, these tools require more time and cost with less accuracy.

Apache Hadoop is best Platform for Big Data processing. Hadoop is open source Java Platform that contains thousands of clusters that is used for parallel processing and execution of Big Data. Its main components are Pig, HBase, Hive, HDFS (Hadoop Distributed File System), MapReduce and Apache Spark Framework. Pig is High level language that is used for scripts. It includes load store operators and provides users the capability of creating own built-in-functions (extensible). HBase is used for automatic sharding and sparse data processing by replacing RDBMS (Relational Database Management System). Hive is not used for real time processing but it is used for large analytics and efficient query processing with the help of meta-store unit. HDFS is file system that is developed for processing and execution of large files in database that is created by Hadoop components. Its two units are data node and name node. MapReduce is designed for parallel execution and processing of large datasets in Hadoop Platform. Apache Spark is framework especially designed for Analytics by using the Languages Java, Python, C and Scala. Its main components are caching, action and transformation.

Many Bioinformatics Algorithms are implemented in Scala language for Apache Spark Framework. Scala is functional, statically typed and object oriented language. It is better for concurrent processing. Its main features are traits, closures and functions that are used for processing of multiple Genome Sequencing Algorithms. Scala mostly works like C++ language.

Scala consists of Arrays, Loops, Strings, Classes, Objects, collections, Pattern Matching and Extractors. All of these structures and statements are used for Bioinformatics Algorithmic comparison by Scala in Spark Framework. Scala also contains many Built-in-Methods, Libraries and Functions that are very useful for designing Bioinformatics Algorithms. Scala language plays an imperative role in Bioinformatics Applications.

Genome Sequencing, Motif Finding, Pairwise Alignment and Multiple Alignment are main features for Bioinformatics. Scala language is very important for these Algorithms. In Genome and Multiple Sequencing, a lot of algorithms are used for handling Biological Sequences. These Algorithms are implemented in Scala language. In Apache Spark, Motif Finding Algorithms are implemented using Scala language. In Pairwise Alignment, Scala language is very significant for pattern Matching.

Spark provides the facility of Scala shell for the implementation of these Bioinformatics Algorithms. Primitive Types and anonymous functions in Scala perform well for managing arrangements of Multiple Sequences. Anonymous functions are used in transformations, actions and loading files for Analytics of Bioinformatics datasets in Apache Spark Framework. Shared variables and key-value pairs are used in Hadoop using Scala language for Bioinformatics Algorithms.

For implementing Bioinformatics Algorithms in Scala language on Hadoop Platform, datasets are stored in specific format. Different storage formats are used for different Algorithms on Hadoop and Spark Platform for example, Fasta, Fastq, CSV, ADAM, BAM (Binary Alignment Map)/ SAM (Sequence Alignment Map) and ADAM.

The objectives of this study are:

- To explore the Supported Languages and Supported Platforms for Genome Sequencing, Motif Finding, Pairwise Alignment and Multiple Alignment Algorithms
- To analyze the need for Scala language for the implementation of Bioinformatics Algorithms on Hadoop Platform
- To explore the Scala Language used in existing Bioinformatics tools

The rest of the paper is organized as follows: Section II explains the related work in this field. Section III describes the tools for Bioinformatics. Section IV represents Role of Scala Implementations in Bioinformatics.

II. Related Work

Ali et al. [1] have explained study in which many Machine Learning classification and clustering Algorithms are implemented in Hadoop MapReduce and Apache Spark using Scala language. They also describe the Performance comparison of different Machine Learning Techniques and Algorithms in the perspective of Hadoop and Spark. It illustrates further research ideas in his paper in which Machine Learning Techniques and Algorithms are implemented in Hadoop and Spark Framework. Sarwar et al. [2] have proposed review study about Bioinformatics Tools. They demonstrate the implementations of Tools for Alignment Viewers, Database Search and Genomic Analysis on Hadoop and Apache Spark Framework using Scala language. It also describes further research domains for the implementation of Bioinformatics Tools on Hadoop and Apache Spark using various languages such as Java, Scala and Python.

SeqPig is a library and tool for Analysis and query sequencing data with scalability [3]. It uses the Hadoop engine, Apache Pig, that automatically parallelizes and distributes tasks that are translated into sequence of MapReduce jobs. It provides extension mechanism for library functions supported by languages (Python, Java and JavaScript) and also provides import and export functions for file format such as Fastq, Qseq, FASTA SAM and BAM. It allows the user to load and export sequencing data. SeqPig provides five read statistics. (a) average base quality read; (b) length of reads; (c) base by position inside the read; (d) GC content of read. Finally combined with single script, it is also used for ad-hoc Analysis but SparkSeq is the best option for ad-hoc analysis.

Wiewiorka et al. [4] have launched bioinformatics tool used to build genome pipeline in Scala and for RNA and DNA sequence analysis. The purpose of this work was to determine scalability and very fast performance by analysis of large datasets such as protein, genome and DNA. A new MapReduce model has been developed for parallel and distributed execution in Spark. Data cannot be stored in HDFS without BAM library (for direct access data and support formats). After data storage in Hadoop, Spark queries applied to sequencing datasets and data is analyzed.

Nordberg et al. [5] proposed the BioPig, used for analysis of large sequencing datasets in the perspective of Scalability (scale with data size), Programmability (reduced development time) and portability (without modification Hadoop). To evaluate these three perspectives, Kmer application was implemented to check its performance and compare with other methods. BioPig uses methods (pigKmer, pigDuster and pigDereplicator). Dataset size for Biopig ranges from 100 MB to 500 GB. Biopig is same as SeqPig in such a way that both use Hadoop and Pig environment and same functions (import and export) and similar run time performance. Only difference is that BioPig includes many Kmer applications and wrapper for BLAST that the SeqPig does not have. The limitation of BioPig is the startup latency of Hadoop. This problem is solved by Spark.

Sun et al. [6] presented the Mapping of long sequence by Bwasw-cloud algorithm with the help of Hadoop MapReduce implementation. Many single processor algorithms like BLAST, SOAP and MAQ are struggling for quick reads. Many multiprocessor algorithms perform much better work like BlastReduce and short reads but some problems occur as its performance and expense for equipment. These problems are decreased by Bwasw-cloud algorithm. This algorithm contains three phases (Map, Shuffle and Reduce) by using seed-and-extend technique and sequence alignment functions are mostly implemented in Map phase. The scaling is measured by length of reads, different mismatches and different number of reference chunks, whereas performance is measured as the speedup over this algorithm.

Taylor et al. [7] focused the next-generation sequencing data and its use in bioinformatics field. Hadoop and MapReduce play an important role in NGS. In this work, he has discussed some terminologies such as Hadoop, MapReduce, HBase, Hive, pig and Mahout then their role in bioinformatics field such as CloudBurst software same as BlastReduce (for NGS short read mapping into reference genome), Bowtie crossbow (for genome re-sequencing analysis), Contrial (for assembly DNA short reads without reference genome), R/Bioconductor (for calculating different gene expression in large RNA-seq dataset). Hadoop and HBase also used for Biodoop tool that consist of three algorithms (BLAST, GSEA and GRAMMAR). Hadoop also used for multiple sequence alignment.

Srinivasa et al. [8] have proposed a technique to classify sequences with the help of Distance matrix formula (m*m) and to understand the relationship among different species during evolution using MapReduce model by dividing the sequences into blocks. Dynamic algorithms Needleman-Wunsch and Smith-waterman are limited to number and size of sequence. So, new MapReduce model developed to reduce these limitations. The input format is FASTA format and output in the custom type. It includes three MapReduce jobs: (a) Data preprocessing (b) Cartesian product (c) Sequence alignment. After these three phases, hierarchical clustering is performed by UPGMA (to produce rooted trees). Due to scalability of Hadoop framework, the proposed method for Phylogenetic is suited for large scale problems.

III. Tools for Bioinformatics

There are several Bioinformatics tools those are used for the analysis of small and large datasets. Every tool performs specific function. Different tools are used for sequence analysis, motif finding, database search and genome analysis. These tools require the data to be stored in a specific format for any kind of analysis. These tools are built using different programming languages. It is important to know the specific language in order to customize the tools. The skills in a programming language are more helpful when extending these tools for Hadoop MapReduce or Apache Spark framework.

A. *Motif Finding Tools*

Sequence motifs are repeated patterns that are of biological significance. Many tools are available for motif finding in the nucleotide or protein sequence. These tools are also implemented using different programming languages like C, C++, Java, Perl, FORTRAN, Python, and R. a list of the motif finding tools is presented in TABLE I.

Like the alignment viewer and genomics Analysis, the motif finding tools also implemented in Apache spark and Hadoop MapReduce Framework for the experimentation of Big Data analysis. PMS and BLOCKS are implemented in a Hadoop MapReduce Framework for the Big Data analysis.

TABLE I. Motif Finding Tools

Name	**Sequence Type**	**Language**	**Data Format**	**MapReduce**	**Spark**
PMS [9]	Protein or nucleotide sequence	*Perl, Python, Java, C++*	Fasta	YES [10]	NO
FMM [11]	Nucleotide sequence	*Python, Java*	Fasta	NO	NO
BLOCKS	Protein or nucleotide sequence	*Perl, Python, Java*	Fasta	YES [12]	NO
eMOTIF	Protein or nucleotide sequence	Java	Fasta	NO	NO
Gibbs motif sampler [13]	Protein or nucleotide sequence	C, C++, Java, and Fortran, Python, R [3]	Fasta	NO	NO
HMMTOP [14]	Protein sequence	Perl, Python, C or Fortran	Fasta	NO	NO
I-sites [15]	Protein sequence	Python, C++	Fasta	NO	NO
JCoils	Protein sequence	C++	Fasta	NO	NO
MEME/MAST [16]	Protein or nucleotide sequence	Ruby, Python	Fasta	NO	NO
CUDA-MEME [17]	Protein or nucleotide sequence	*Python, Perl,* Fortran, *Java, Ruby* C, C++	Fasta	NO	NO
MERCI	Protein or nucleotide sequence	C, C++	Fasta	NO	NO

B. *Multiple Sequence Alignment*

These tools are used for the alignment of more than two nucleotide or protein sequences. These tools are helpful in finding the homology and evolutionary relationships between the studied sequences. A number of multiple sequence alignment tools are developed using Ruby, C, C++ and Python.

ABA, ALE, AMAP, anon, BAli-Phy are implemented in Ruby, C, Python and C++. Multiple sequence alignment tools support different format of data for storage and alignment purpose of protein and nucleotide. ABA, ALE, AMAP, anon, BAli-Phy tools have the different data format like Fasta GenBank, EMBL, GDBM, PHYLIP, MFA. With the growing technologies in Bioinformatics, the tools of Multiple Sequence Alignment are also implemented in Modern technology like Hadoop MapReduce and Apache Spark. MSA, SAGA MSAProbs are tools of Multiple Sequence Alignment category that are implemented in Hadoop MapReduce and Apache Spark.

TABLE II presents the available multiple sequence alignment tools.

TABLE II. MULTIPLE SEQUENCE ALIGNMENT TOOLS

Name	Sequence Type	Language	Data Format	MapReduce	Spark
ABA [18]	Protein sequence	Ruby	Fasta	NO	NO
ALE	Nucleotides	C Python	GenBank, EMBL Fasta GDBM Phylip,	NO	NO
AMAP	Protein and Nucleotides sequence	Python	MFA Fasta	NO	NO
Anon	Nucleotides	Python	-	NO	NO
BAli-Phy	Protein and Nucleotides sequence	C++	Fasta	NO	NO
Base-By-Base [19]	Protein and Nucleotides sequence	Java	Fasta GenBank	NO	NO
CHAOS/DIALIGN	Protein and Nucleotides sequence	Java	Fasta	NO	NO
ClustalW	Protein and Nucleotides sequence	C++	Fasta	NO	NO
CodonCode Aligner	Nucleotides	C++	Fasta Fastq, Sam, GenBank, or EMBL	NO	NO
Compass [20]	Protein sequence	C, C++, Java Python	Fasta	NO	NO
DECIPHER	Protein and Nucleotides sequence	R	Fasta Fastq , QSEQ, RAW, Miro, and Seq	NO	NO
DIALIGN-TX and DIALIGN-T	Protein and Nucleotides sequence	C	Fasta	NO	NO
DNA Baser Sequence Assembler	Nucleotides	Java	SCF, ABI, Fasta SEQ, TXT, GBK	NO	NO
DNASTAR Lasergene Molecular Biology Suite	Protein and Nucleotides sequence	C, C++ Python	EMBL, GenBank	NO	NO
DNA Alignment	Protein and Nucleotides sequence	Python Perl Javascript	Fasta	NO	NO
EDNA [21]	Nucleotides	Java	GeneMappe	NO	NO
FSA	Protein and Nucleotides sequence	C++	Fasta	NO	NO
Geneious	Protein and Nucleotides sequence	C++	Fasta Genbank	NO	NO
Kalign	Protein and Nucleotides sequence	C	Fasta GCG, EMBL, GenBank, PIR,NBRF, Phylip, Swiss-Prot	NO	NO
MAFFT	Protein and Nucleotides sequence	C	Fasta	NO	NO
MARNA	RNA sequence	C++	Fasta	NO	NO
MAVID [22]	Protein and Nucleotides sequence	C++	Fasta	NO	NO
MSA	Protein and Nucleotides sequence	C	Genepop , Msvar, Structure, Arlequin, Migrate, IM-format	YES [23]	NO
MSAProbs	Protein sequence	C++ bioPerl	Fasta	YES [23]	NO
MULTALIN	Protein and Nucleotides sequence	C	MultAlin, Fasta , GenBank , EMBL, SwissProt	NO	NO
Multi-LAGAN [24]	Protein and Nucleotides sequence	C C++	Fasta	NO	NO
MUSCLE	Protein and Nucleotides sequence	C++	Fasta	NO	NO
Opal	Protein and Nucleotides sequence	Java	data maNOger file (*.odm)	NO	NO
Pecan [25]	DNA sequence	Python	Fasta	NO	NO
Phylo	Nucleotides	R, Javascript	Fasta	NO	NO
PMFastR	RNA sequence	C++	Fasta	NO	NO
Praline	Protein sequence	Ruby Javascript	Fasta or PIR	NO	NO
PicXAA [26]	Protein and Nucleotides sequence	C++	Fasta	NO	NO

POA	Protein sequence	C	Fasta	NO	NO
Probalign	Protein sequence	C++	Fasta	NO	NO
ProbCons	Protein sequence	C++	Fasta	NO	NO
PROMALS3D	Protein sequence	Python	Fasta	NO	NO
PRRN/PRRP [27]	Protein sequence	Ruby	Fasta	NO	NO
PSAlign	Protein and Nucleotides sequence	C	Fasta	NO	NO
RevTrans	DNA or Protein	Python	ASTA, MSF and ALN	NO	NO
SAGA [28]	Protein sequence	C	Fasta	YES [29]	NO
SAM	Protein sequence	Perl, C	Fasta	NO	NO
Se-AL	Protein and Nucleotides sequence	Java	Nexus, Phylip, MEGA, NBRF, Fasta GDE and GDE 97	NO	NO
StatAlign [30]	Protein and Nucleotides sequence	Java	Fasta	NO	NO
Stemloc	RNA sequence Alignment	-	Fasta	NO	NO
UGENE	Protein and Nucleotides sequence	C++, Qt	Fasta, GenBank , EMBL , GFF	NO	NO
VectorFriends	Protein and Nucleotides sequence	Assembly	EMBL, Fasta Nexus	NO	NO
GLProbs [31]	Protein sequence	C++	Fasta	NO	NO
T-Coffee	Protein and Nucleotides sequence	C, biopython C++, Perl and python	Fasta , PIR	NO	NO

C. Pairwise Alignment

These tools are used for the identification of similarity regions between two biological sequences that can indicate functional, structural or evolutionary relationships. Pairwise Alignment tools are also implemented in different programming languages. ACANA is implemented in C++, AlignMe in Python and Perl, Bioconductor in PHP, Perl and Java, BioPerldpAlign in Perl, BLASTZ, LASTZ in C and CUDAlign is implemented in C++. A list of the available pairwise alignment tools is presented in TABLE III.

Pairwise alignment tools also support different data formats for the storage and analysis of biological data. These formats include Fasta Fastq, BAM, gtf, bed, wig, nib, hsx, GenBank, Raw DNA file formats, and Primers (.csv). Some of the tools also support Hadoop MapReduce and Apache Spark. Matcher, JAligner, Genome Compiler, Bioconductor, BioPerldpAlign are tools that are implemented for Big Data Platforms.

TABLE III. PAIRWISE ALIGNMENT TOOLS

Name	**Sequence Type**	**Language**	**Data Format**	**MapReduce**	**Spark**
ACANA [32]	Protein or nucleotide sequence	C++	Fasta	NO	NO
AlignMe	Protein sequence	Python,Perl	Fasta	NO	NO
Bioconductor	Protein or nucleotide sequence	PHP, Perl Java	Fasta fastq, BAM, gtf, bed, and wig	YES [78]	NO
BioPerl [33]	Protein or nucleotide sequence	Perl	Fasta	YES [79]	NO
BLASTZ,LASTZ	Nucleotides	C, C++	Fasta fastq, nib, 2bit or hsx	NO	NO
CUDAlign	Nucleotides	C++	Fasta	NO	NO
DNADot	Nucleotides	Java	Fasta	NO	NO
DNASTAR Lasergene Molecular Biology Suite	Protein or nucleotide sequence	Java	GenBank	NO	NO
DOTLET	Protein or nucleotide sequence	Java	Fasta	NO	NO
FEAST [34]	Nucleotides	Java	Genbank	NO	NO
Genome Compiler [35]	Nucleotides	*C, Perl, PHP, Java,* ruby Python, Perl	GenBank, Fasta	YES [80]	YES [81]
G-PAS	Protein or nucleotide sequence	C++	Fasta	NO	NO
GapMis	Protein or nucleotide sequence	C	Fasta	NO	NO
GGSEARCH, GLSEARCH	Protein sequence	C, C++	Fasta	NO	NO
JAligner [36]	Protein or nucleotide sequence	Java	Fasta	YES [82]	NO

K*Sync	Protein sequence	Java	Fasta	NO	NO
LALIGN	Protein or nucleotide sequence	Python	Fasta	NO	NO
NW-align [37]	Protein sequence Alignment	Java	Fasta PDB format	NO	NO
mAlign	Nucleotides	Java, C	Genbank Fasta	NO	NO
Matcher	Protein or nucleotide sequence	C, C++	Fasta msf, trace, srs	YES [83]	YES [84]
MCALIGN2 [38]	DNA sequence Alignment	C++	Fasta	NO	NO
MUMmer	Nucleotides	-	Fasta delta	NO	NO
Needle	Protein or nucleotide sequence	C, C++, Python	Fasta msf, clustal, mega, meganon, nexus,,nexus	NO	NO
Ngila [39]	Protein or nucleotide sequence	C++	Fasta	NO	NO
NW	Protein or nucleotide sequence	C, C++, Python	Fasta	NO	NO
Parasail	Protein or nucleotide sequence	C, C++, Python	Fasta Fastq	NO	NO
Path [40]	Protein sequence Alignment	Java	Fasta	NO	NO
PatternHunter	Nucleotides	Java	Genbank Fasta	NO	NO
ProbA (also propA)	Protein or nucleotide sequence	C	Fasta	NO	NO
PyMOL	Protein sequence Alignment	C, C++	Fasta Genbank	NO	NO
REPuter [41]	Nucleotides	Json web service	Fasta Genbank	NO	NO
SABERTOOTH	Protein sequence Alignment	Java	FastaGenbank, EMBL, SWISSPROT	NO	NO
Satsuma	DNA sequence	C++	Fasta	NO	NO
SEQALN [42]	Protein or nucleotide sequence	-	genbank, newat, Fasta, pir, swissprot	NO	NO
SIM, GAP, NOP, LAP	Protein or nucleotide sequence	C/C++/Python	Fasta	NO	NO
SIM	Protein or nucleotide sequence	C/C++/Python	Fasta	NO	NO
SPA: Super pairwise alignment	Nucleotides	C++	Fasta Genbank	NO	NO
SSEARCH	Protein sequence	C#, Java, Perl	Fasta	NO	NO
Sequences Studio [43]	Generic Sequence	Java	Fasta	NO	NO
SWIFT suit	DNA sequence	Swift	Fasta	NO	NO
Stretcher	Protein or nucleotide sequence	Ruby	Fasta Genbank	NO	NO
ss	Nucleotides	R	Embl, Imgt Refseqn Genbank	NO	NO
UGENE	Protein or nucleotide sequence	C++	FASTA GenBank, EMBL, GFF	NO	NO
Water [44]	Protein or nucleotide sequence	R	Fasta	NO	NO
WordMatch	Protein or nucleotide sequence	R	Fasta msf, clustal, mega, meganonexus	NO	NO
YASS [45]	Nucleotides	C	Fasta Axt	NO	NO

IV. Importance of Scala Implementations

A lot of programming languages are being used for the implementation of Bioinformatics Algorithms on Hadoop Platform and Apache Spark. Most of Bioinformatics tools are implemented using Java, Python, C++, Perl, FORTRAN, R, Ruby, C, Bioperl, Assembly, JavaScript, PHP and Swift languages. Some Algorithms are used in Multiple Sequence Alignment, Pairwise Alignment and Motif Finding. These Algorithms are implemented by using Hadoop and Apache Spark framework. Many languages are used to implement these Bioinformatics Algorithms. Most commonly used languages are Java, Python and Scala.

Our goal is to use best language for the implementation of Bioinformatics Algorithms. Scala language is superlative language in Hadoop Platform and Apache Spark for the implementation of Bioinformatics Algorithms. By using Scala language for Bioinformatics Algorithms, we will achieve better Performance, Scalability and Accuracy. This language plays imperative role in all benchmarks. When we implement Bioinformatics Algorithms in Spark Framework, Scala language give better results. Closure, Traits, Pattern Matching and Functions are main features of Scala language.

Some Motif Finding tools such as PMS, FMM, BLOCS, eMOTIF, Gibbs motif sampler, HMMTOP, JCoils, MEME/MAST, CUDA-MEME and MERCI are available in Bioinformatics. Algorithms in these tools are not implemented in Spark using Scala language. We can use Scala language for the implementation of these Motif Finding Bioinformatics Algorithms to attain better outcomes. Scala is state of the art language that associates Object Oriented and Functional programming concepts.

Most of tools such as ABA, ALE, AMAP, Anon, Bali-Phy, Base-By-Base, CHAOS, ClustalW, CodonCode Aligner, Compass, DECIPHER, DNA Alignment, Geneious, Kalign, EDNA, FSA, MAFFT, MARNA, MAVID, MSA, MUSCLE, Opal, Pecan, Phylo, Praline, POA, PicXAA, ProbCons, PSAlign, SAGA, SAM, Se-AL, StemAlign, UGENE and VectorFriends are available for Multiple Sequence Alignment in Bioinformatics. Some Pairwise Alignment tools such as ACANA, AlignMe, Bioconductor, BioPerl, BLASTZ, CUDAlign, DNADot, DOTLET, FEAST, Genome Compiler, G-PAS, GapMis, JAligner, K*Sync, LALIGN, NW-Align, Matcher, MUMmer, Needle, Ngila, NW, Parasail, Path, ProbA, PyMOL, REPuter, Satsuma, SIM, GAP, NOP, LAP, SIM, SSEARCH, Sequences Studio, SWIFT suit, Stretcher, SPA, ss, UGENE, Water and YASS are available for Bioinformatics. Algorithms in these tools are not implemented in Spark using Scala language. We can use Scala language for the implementation of these Multiple Sequence Alignment and Pairwise Alignment Bioinformatics Algorithms to attain better outcomes.

Many Bioinformatics Algorithms are based on Greedy and Dynamic Programming paradigm. Some Bioinformatic sequences are Map/Align with Local, Global, Multiple and Pairwise method. Nussinov-Algorithm and Viterbi-Algorithm also require Scala language for their implementation. SCABIO is the best framework for Bioinformatics Algorithms in Scala language. It includes many built-in-methods and libraries that are helpful for Scala implementation. It also provides Greedy and Dynamic Programming approach for Bioinformatic sequences. We can use SCABIO for Global, Local, Multiple and Pairwise Alignment. Pattern Matching is best performed with the help of SCABIO because SCABIO includes Scala language implementation concepts.

V. Conclusion

Keeping in view the data analysis demands in Bioinformatics, Big Data Platforms and tools are an obvious choice. Among these platforms, Spark is most efficient platform for rapid analysis of large data sets. Spark itself is implemented in Scala languages and supports programs in Java, Scala and Python. Majority of the tools in bioinformatics are not designed for Big Data Platforms. As discussed in previous sections, most of the Multiple Alignment tools, Pairwise Alignment tools and Motif Finding tools still need to be enhanced for use on Big Data Platforms like Hadoop and Spark. So, there is need of time to implement bioinformatics tools on Big Data Platforms. Several languages are available for implementation of bioinformatics tools like Java, C, Perl, Python and Scala. Among these languages, Scala is a good choice especially for Spark Implementations. It provides structures and constructs that are suitable for Bioinformatics applications. It provides support for dynamics programming and pattern matching. It can provide efficient implementations of machine learning algorithms. We recommend that Scala must be used for future implementations of Bioinformatics tools on Big Data Platforms.

References

[1] M. U. Ali, S. Ahmad and J. Ferzund, "Harnessing the Potential of Machine Learning for Bioinformatics using Big Data Tools," International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 10, pp. 668-675, 2016.

[2] M. A. Sarwar, A. Rehman and J. Ferzund, "Database Search, Alignment Viewer and Genomics Analysis Tools: Big Data for Bioinformatics," International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 12, 2016.

[3] S. Andre, P. Luca, N. Matti and K. Aleksi, "SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop," Oxford.

[4] S. Oehmen, "ScalaBLAST 2.0: rapid and robust BLAST calculations on multiprocessor systems," Oxford.

[5] N. Henrik, B. Karan, W. Kai and W. Zhong, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data," oxford, September 10, 2013.

[6] S. Mingming, Z. Xuehai and Y. Feng, "Bwasw-Cloud: Efficient sequence alignment algorithm for two big data with MapReduce," in Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference, 2014.

[7] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," BioMed Central, 2010.

[8] G. Siddesh, K. Srinivasa and M. Ishank, "Phylogenetic Analysis Using MapReduce Programming Model," in Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2015 IEEE International, 2015.

[9] "http://motifsearch.com/," [Online]. Available: http://motifsearch.com/.

[10] "A MapReduce-based Algorithm for Motif Search," ResearchGate.

[11] "Learn Motifs from Unaligned Sequences," [Online]. Available: https://genie.weizmann.ac.il/pubs/fmm08/fmm08_learn_unalign.html.

[12] Y. Liu, X. Jiang, H. Chen, J. Ma and X. Zhang, "MapReduce-Based Pattern Finding Algorithm Applied in Motif Detection for Prescription Compatibility Network," Springer Link.

[13] "Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences.," PubMed.

[14] "The HMMTOP server," [Online]. Available: http://www.enzim.hu/hmmtop/html/document.html.

[15] "I-sites Libraries 2008," [Online]. Available: http://www.bioinfo.rpi.edu/bystrc/Isites2/.

[16] [Online]. Available: https://en.wikipedia.org/wiki/Multiple_EM_for_Motif_Elicitation.

[17] "cuda-meme," [Online]. Available: https://sites.google.com/site/yongchaosoftware/Home/cuda-meme.

[18] R. Benjamin, D. Zhi, H. Tang and P. Pevzner, "A novel method for multiple alignment of sequences with repeated and shuffled elements," PubMed Central.

[19] "Virology.ca Tools," [Online]. Available: http://athena.bioc.uvic.ca/virology-ca-tools/base-by-base/.

[20] G. N. Sadreyev R, "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance.," PubMed .

[21] [Online]. Available: https://sourceforge.net/projects/msa-edna/.

[22] N. B. Pachter and Lior, "MAVID multiple alignment server," Oxford.

[23] Jurate, O. D. Aisling and D. S. Roy, "An Overview of Multiple Sequence Alignments and Cloud," ISRN Biomathematics, 2013.

[24] Brudno M et al. "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.," PubMed.

[25] [Online]. Available: https://github.com/benedictpaten/pecan/blob/master/doc/pecan/README_PECAN.txt.

[26] S. Sahraeian and B. Yoon, "PicXAA: a probabilistic scheme for finding the maximum expected accuracy alignment of multiple biological sequences.," PubMed.

[27] "prrn," [Online]. Available: http://www.genome.jp/tools/prrn/prrn_help.html.

[28] "SAGA HOME PAGE," [Online]. Available: http://www.tcoffee.org/Projects/saga/.

[29] M. Chris and M. Michael, "Programming Abstractions for Data Intensive Computing on Clouds and Grids," CCGRID '09 Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid.

[30] "An Extendable Software Package for Joint Bayesian Estimation of Alignments and Evolutionary Trees," [Online]. Available: http://statalign.github.io/.

[31] Y. Yongtao, W.-l. C. David and W. Yadong, "GLProbs: Aligning Multiple Sequences Adaptively," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014.

[32] W. Huang, D. Umbach and L. Li, "Accurate anchoring alignment of divergent sequences.," PubMed.

[33] "BioperlOverview," [Online]. Available: http://www.ebi.ac.uk/~lehvasla/bioperl/BioperlOverview.html.

[34] [Online]. Available: http://monod.uwaterloo.ca/feast/..

[35] "Genome Compiler," [Online]. Available: http://www.genomecompiler.com/about-genome-compiler/.

[36] [Online]. Available: http://jaligner.sourceforge.net/.

[37] [Online]. Available: http://zhanglab.ccmb.med.umich.edu/NW-align/.

[38] W. Jun, D. K. Peter and J. Toby, "MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution," BMC Bioinformatics.

[39] [Online]. Available: http://scit.us/projects/ngila/ .

[40] [Online]. Available: http://bioinfo.lifl.fr/path/path.php.

[41] [Online]. Available: http://bibiserv.techfak.uni-bielefeld.de/reputer/.

[42] [Online]. Available: http://thegrantlab.org/bio3d/html/seqaln.html.

[43] [Online]. Available: http://www.bioinformatics.org/sStu/doc/index.html.

[44] [Online]. Available: http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/water.html.

[45] [Online]. Available: https://en.wikipedia.org/wiki/Yass_(software).

Analytical Review on Test Cases Prioritization Techniques: An Empirical Study

Zainab Sultan
Department of Software Engineering
Bahria University Islamabad, Pakistan

Shahid Nazir Bhatti
Department of Software Engineering
Bahria University Islamabad, Pakistan

Rabiya Abbas
Department of Software Engineering
Bahria University Islamabad, Pakistan

S. Asim Ali Shah
Department of Electrical Engineering
Bahria University Islamabad, Pakistan

***Abstract*—For conclusively predicting the quality of any software system, software testing plays an important but a vital role. For finding faults early and to observe failures (anomalies) before implementation stage, software testing is done and if bugs (defects) are detected then software is passed through maintenance phase. The success and failure of a software project is often attributed to the development methodology used. It is also observed that in many scenarios, the software engineering methods are not implemented in their true spirit. Moreover, many of the development methodologies don't cater the change very well, because they follow a predefined development path which allows very less deviation. In software testing, regression testing is the important type of software testing. When any change made on the software then regression testing is done to check that it doesn't influence other parts of software. In regression testing, test cases are prioritized in order to reuse new test cases and existing test cases. Test case prioritization is done by using different techniques. This paper presents a review of different test case prioritization techniques.**

Keywords*—Agile Software Engineering (ASE); Testing; Regression Testing; Test Suit Reduction; Test Case Generation; Test minimization; Test Case Prioritization Technique

I. Introduction

With quality defined as "meeting requirements, testing defines the quality as "fulfillment of the requirement specification", thus testing gives a good idea of the quality level. This leads to main objective of testing i.e. "Testing reduces the level of uncertainty about the quality of an software system. Software testing is the most significant step of software development life cycle. The testing involves the programs or an application's implementation having aim to discover software bugs and faults. There are different types of testing that software tester adopt according to their requirements such as Mutation Testing, Regression Testing, stress testing, security testing, load testing, black box testing, white Box Testing. According to testing type, the tester creates the number of test cases is called Test Suite. In testing process duration, tester finalizes test cases, implement on software according to developed test cases and then verify and check the results come by those executions. Regression testing is a testing technique which is applied on the altered application using pre-defined sets of Test cases.

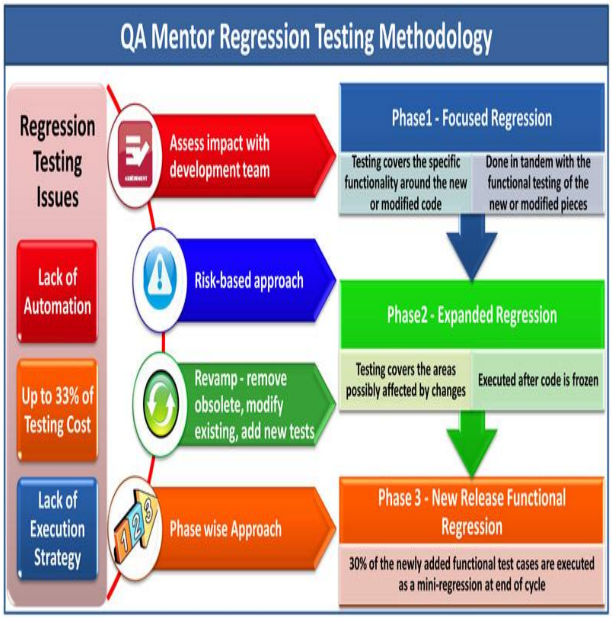

QA Mentor Regression Testing Methodology
Regression Testing Issues
Lack of Automation
Up to 33% of Testing Cost
Lack of Execution Strategy
Assess impact with development team
Risk-based approach
Revamp - remove obsolete, modify existing, add new tests
Phase wise Approach
Phase1 - Focused Regression
Testing covers the specific functionality around the new or modified code
Done in tandem with the functional testing of the new or modified pieces
Phase2 - Expanded Regression
Testing covers the areas possibly affected by changes
Executed after code is frozen
Phase 3 - New Release Functional Regression
30% of the newly added functional test cases are executed as a mini-regression at end of cycle

Fig. 1. Regression testing methodology

When an application is first time tested, test suite is constructed to enhance its functionality. Tester preserve test suite for further use. As changes are made in system, then these pre-defined test suites are applied by testers so that it can be ensured that no new bugs are introduced in the code that have been tested. If changes occur in system, then re executing every test for each module after change is really inapplicable and illogical. Moreover, it is much costly approach to execute all test cases once changes made. So to decrease the regression testing cost and to mold it in more profitable form, "test case prioritization" concept was introduced by the researchers. In test case prioritization, all test cases are arranged in an order to magnify some equitable behavior. To establish the priorities of test cases certain factors depending upon the requirement are analyzed and selected and then preference is allocated to test cases. Test case prioritization delivers a path to lineup and executes test cases, which has maximum priority in order to detect earlier faults. Different test case techniques of prioritization are reviewed in this review paper. For researchers, it can provide help to find which technique is best suitable for which scenario.

II. Regression Testing Approaches

Here are some approaches for regression testing [1].

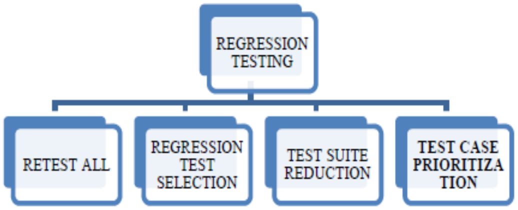


Fig. 2. Approaches of Regression Testing

1) *Retest All*

For regression testing, it is the most genuine Technique. During this approach, simply in test suite all test cases execute.

2) *Regression Test Selection*

It deals with the problem arising, from test suite for adopting a subset of test case. Then chosen test cases are executed to verify and test modifications occur in program.

3) *Test Suite Reduction*

This process containing two parts: First is, identification of relevant or superfluous test cases, second one is, exclude those test cases.

4) *Test Case Prioritization*

It includes with the realization of idealized sorting of test cases. That ordering must increase the fascinating properties like early detection of faults and no of fault detection and minimizes the cost factor.

III. Test Case Life Cycle

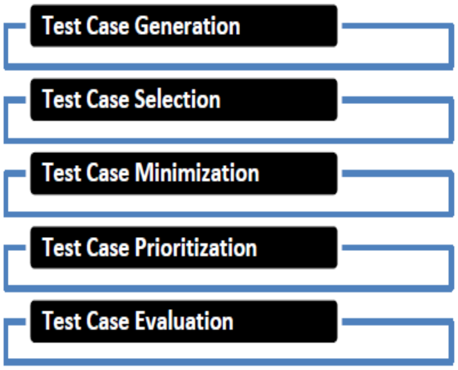


Fig. 3. TCLC test case life cycle

Figure 3 presents Life cycle of test case (TCLC). Testing of software is important, significant and profitable process in SDLC (software development life cycle). In testing process, on a set of different test cases tester executes program, and compare actual outcome with expected outcome. From different software artifacts like requirements specifications and designs test cases are mostly extracted. Testers study requirement document first to understand the requirements and specifications, when they understood requirements they start preparing test cases. Using tools they are automatically generated. Different techniques are used for test cases generation. Various phases of test case life cycle are, "Test case generation, test case selection, test case minimization, test case prioritization and evaluation". Test case generation is the process of generating test suites for a particular system. Some methods of test case generation relays on application, like test case generation for object oriented application, web application, UML applications, applications based on evolutionary and genetic algorithms, structured based systems, and many others. Test cases are categorized into five classes as "reusable, retest able, obsolete, structural, new specification and new structural test cases". In test case selection from a test suite choosing test cases in software testing process for reduction of time, Effort and cost. It is just like to test case minimization technique. The Test suite minimization approach, relay on metrics such as from a single version measurement of coverage of the program under test. Diversity among these two approaches builds upon the modifications occurred in SUT. According to the changes made among preceding and ongoing version of the SUT, the test cases are chosen. In Test case prioritization from multiple test suites of software test cases are conscripted, ranked and arranged. To rank and organize the test cases there are a lot of techniques. Some priority is accredited to each test case; however when multiple test cases are assigned the same priority or weights problem originates sometimes. In test case evaluation, to test the software, test cases are appraised to determine the suitable test cases. To evaluate the test cases, perform: (1) Prepare experiment data, (2) Run the test suites prioritization method, (3) Evaluate results.

IV. Test Case Prioritization Vs Random Test Cases

In a research paper [3], researchers took a bank application project and compared prioritized and random test cases. They found, more faults can be identified if test cases will be prioritized. Results are displayed in figure 4.

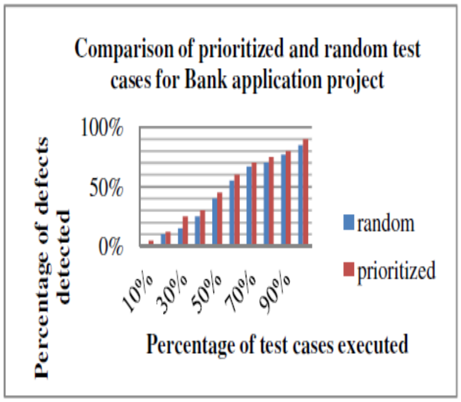


Fig. 4. Prioritized and Random comparison

Some attributes are taken and evaluated them and compared either proposed techniques by different authors are better or random ordering.

Attributes are: Comparison based on 1) size of test cases 2) Time taken by test cases 3) Effort taken by test cases 4) Cost taken by test cases 5) Efficiency 6) More defects found by test cases 7) Can be used in other projects.

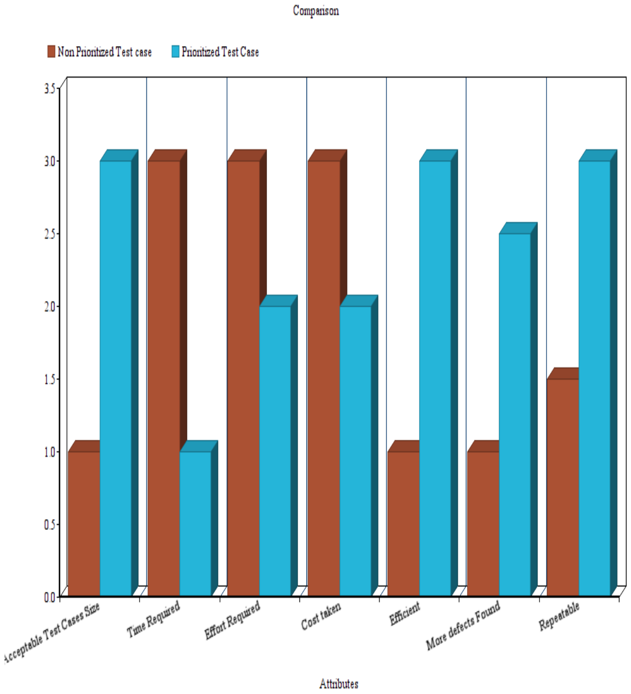


Fig. 5. Comparison b/w Random and prioritized test cases on different Attributes

Hence the test cases which are prioritized give much exceptional fault discovery than the test cases which are not prioritized. Further using techniques of test case prioritization, diminish the project's time and budget by prioritizing the most significant test cases.

V. Factors for Classification of Test Case Prioritization

Prioritization of test cases relays on following certain factors. In this section, the assorted factors are explained, prioritization depends. [2]

1) *Customer Requirements*
2) *Coverage Based*
3) *Cost effective*
4) *History Based*

1) *Customer Requirements*

In "customer requirements based prioritization techniques", test cases are arranged having focus on customer's requirements which are documented all the while requirements specification collection phase. Key points to prioritize the test cases for this approach are "Assigned property of customer (CP), Complexity of requirement (RC), and volatility of requirement (RV)". Values are accredited to these factors and high value factor illustrates a need for prioritization of test cases.

2) *Coverage Based*

Prioritization technique which depends on coverage, during testing the test case prioritization are on validating and calculating a program's source code that has been executed. [2] Word "coverage" refers during the process of testing the form of code that has coated, it can be "coverage of requirement, coverage of total requirement, and coverage of additional requirement", hence test case has capacity in this approach to test the main code's parts and prioritize them.

3) *Cost effective*

In this approach, test cases are prioritized depending and analyzing cost factor. Cost conceivable as, "requirement gathering cost, regression testing cost, execution and validating cost of test cases, analyses cost and prioritization cost of test cases. Therefore, test cases demanding the diminish cost have maximal value.

4) *History Based*

In "History based", Test cases are scheduled depending upon history of test cases that refers priority relays on test case's prior achievements. The past performances of test cases boost or decline the possibility in order that it will be valuable in testing ongoing session.

VI. Literature Review

This section mentions and summarizes some test case prioritization research papers.

Hema Srikanth and Laurie Williams in their paper [4] introduced a technique in which they target three aspects: on requirements Customer assigned priority (CP), complexity of requirements (RC), Volatility of Requirement (RV). Customers allocate the CP value. Developers assign RC value. They allocated values from 1 – 10 to every factor for analysis. They declared the factors having high values identify the demand of prioritization of test cases. Weight prioritization (WP) indicates significance of verifying requirement at early stage, and it is represented by below equation:

$$WP = \sum_{PF=1}^{n} (PF\ value\ *\ PF\ weight)$$

Here, n denotes absolute no of test cases, WP illustrates weight prioritization, and PF value is allocated to each factor i.e. CP, RC and RV. PF weight is assigned to every factor such as CP, RC and RV. Test cases are organized in corresponding way so that, having higher WP value is performed before others.

Siripong [5] described 4C classification in his research paper. He categorized techniques into 4 bases: Based on "Requirements of customer, Coverage, Cost, and Chronographic history". The Author also defined two methods for test case prioritization. In MTSSP Method: MTSSP was developed to figure out issues originated in multiple test suites allocated to the similar preference. In test suite, on the basis of defect factor test cases are assigned preference. If issue is still not resolved then the test suite is arranged on the basis of timeline aspect. Again if the problem is not fixed then the test suite is prioritized according to budget factor. If this experiment does not succeed then align them with complex factor. If the obstacle remains same then random method is used for test suite prioritization. MTSPM intention is to prioritization of multiple test suites effectively. Test suites are only prioritized by time and cost factors. If the issue does not resolved and test suites are having the similar priority then they are randomly organized. In test suites to prioritize the test cases MTSSP method is used. After MTSSP if the priority remains same, then MTSPM method is applied to prioritize them.

Srivastata [6] recommended test case prioritization according to criteria of increased APFD (Average percentage of fault detected) rate. Author invented a new approach in which test case determine the average no of faults found per minute. Finding that value arranges the test cases in descending order, APFD is determined by equation

$$APFD = 1 - \left(\frac{TF_1 + TF_2 + \ldots\ldots\ldots TF_m}{nm}\right) + \left(\frac{1}{2n}\right)$$

Here, under assessment T is test suites, m represents number of faults found in software or module of software under evaluation, n denotes absolute number of test cases. TFi represents number of test cases that are possessed in test suite T, that reveals faults i. If the previous knowledge of faults is known then APFD can be applied.

Korel et al. [7] implemented an experiment in order to validate and verify efficiency and effectiveness of both simple code based and model based test case prioritization. The Aim of this experiment was to verify these approaches to evaluate the performance of earlier fault detection in the system that has modified. So the result shown that as the execution of model based test cases are very quick as compared to code based test case prioritization so it can bring improvement in earlier detection of fault. That's why for the whole test suites the model based test prioritization in comparison of code based test case prioritization is cheap.

Korel et al. [8] in their paper prioritized the test cases by applying certain model based test prioritization heuristics. It has some problems that selective model based prioritization focus only the number of identified transitions which does not have valuable impact on the improvement of earlier detection of faults. Value based regression test case prioritization [9] is used for the earlier fault detection. For prioritization of test cases, they proposed an algorithm on the basis of 6 factors. Modifications in requirement, Priority of customer, complexity in Implementation, traceability in requirement, time of execution and Impact of fault. The PSO (Particle swarm Optimization) is applied for allocating utility to factors and comparison of factor's values. Maximum value is the highest number, and minimum value is the lowest number. For all the test cases, sum up the values of factors. If the factor values of two test cases match, then it shows by comparing requirement utilities of those test cases the judgement is made. This implies, in terms of time and cost PSO algorithm is much efficient, powerful and useful than greedy algorithm.

Kumar et al. [10] proposed that the prioritization is set on the basis of harshness of faults. The total severity of faults detection (TSFD) is addition of severity frequency of all defects that are exposed in a product is given in below equation. Here n shows total number of faults occur in product.

$$TSFD = \sum_{i=1}^{i=n} SM$$

R.Beena et al. [11] has given an effective and efficient way of choosing and organizing of test cases on the basis of coverage of code. This technique is much significant for decreasing time and budget for regression testing. This approach consists of 3 techniques. "Minimization, selection and prioritization of test cases". TCS algorithm considers for selection of test cases and TCP algorithm considers for prioritization. Test cases are assembled in three categories, Outdated, Required and Surplus. TCCij is matrix that represents the test cases and the statements covered. SDELi is vector that represents the statements deleted in P. SMODi is vector that represents the statements modified in P. These all 3 are the input and output will be the modified matrix TCCij, cluster of test cases, out datedi, surplusi, requiredi. Any statement that cover any test case or many test cases, is considered as out dated. The statements that are modified and are not covered by any statement will be added into surplus cluster and will be removed from TCCij. Remaining test cases are added into required cluster. So the original TCCij will be greater than the required TCCij. .For prioritization, that test case selection output is considered as input of the prioritization algorithm TCPi and output will be TCPi which is vector and consist of test cases to get 100% code coverage. In this algorithm, the statements that are camouflaged by test cases, from new or needed TCCij they are summed up. Choose test cases having maximum value and include it into the TCPi Vector. Remove TCPi Vector from TCCij. Repeat all these steps till all the statements are deleted.

Parakash et al [12] invented a new method for test case prioritization known as "potentially weighted method". This approach prioritizes test cases based on "potential coverage" like coverage of code, function, branch, fault, and path and for criteria weights are assigned. Test cases are given preference on the basis of value of weight, and the weight is from high to low. Every statement in code must be executed at least once; this is the basic purpose of code coverage testing. The criteria value Ticd is determined as

$$T_i^{cd} = \left(\frac{N_{cd}}{M_{cd}}\right) * 10$$

Line of codes that are coated by test case Ti, denoted by Ncd. Any test case Ti that covered Maximal number of codes represented by Mcd. The function coverage testing is very useful as code must be executed at least one time. Function coverage is defined as

$$T_i^{fn} = \left(\frac{N_{fn}}{M_{fn}}\right) * 10$$

Here no of functions are denoted by Nfn that are measured by test case Ti. Test case Ti covered maximum number of functions denoted by Mfn. To increase the capability and performance and for reduction of budget and time this approach is very useful.

Praveen [13] proposed the paper in which average faults per minute are determined; test cases are prioritized on the basis of fault detection rate. For test case prioritization, author invented a new algorithm. Average fault per minute is calculated in this algorithm.

$$AF/m = \frac{F}{Tcost}$$

In the algorithm input is T which is Test suite, identification of number of faults by test case f, and to run each test case cost required is Tcost and and output will come as prioritized test suite. After calculating the fault identified per minute, on the basis of each test case value arrange T in descending order. With the help of APFD analysis has done for prioritized and non-prioritized cases. Author proved with the help of graphs that in the experiment and analysis that test cases which are prioritized are more efficient and useful.

Kavitha et al [14] invented a technique on the basis of rate of fault impact and fault detection to prioritize the test cases. For identification of dangerous faults at earlier unique algorithm is invented. The invented algorithm determines the test case weight age.

$$Tcw = RFT + Fl$$

Tcw denotes test case weightage age.

$$T_{cw} = RFT_i + Fl_i$$

RFTi denotes fault detection rate, average no of faults per minute by test case, is knows as fault detection.

$$RTFi = \left[\left(numberof\ \frac{faults}{time}\right)*10\right]$$

Here, Fli denotes fault impact.

$$FLi = \left(\frac{Si}{\max(s)}\right)*10$$

Where, Si Denotes Test case value.

$$Si = \sum_{j=1}^{t} SV$$

Max(s) denotes high level of severity. The algorithm prioritizes the test cases on the basis on test case weightage. Results have proved, the proposed algorithm is efficient and useful.

Using genetic algorithm [15] invented testing that includes determination of the test cases, which can able to detect bugs in system. The process is tough and time taking. Author proposed a new technique in this paper, for test case prioritization according to their capabilities of discovering bugs. Higher priority is assigned to more similar errors. Low priority is assigned to less similar errors. Through genetic algorithm this order will be achieved. For finding the fittest chromosome like selection, crossover, mutation is applied on chromosomes. Genetic algorithm Steps includes: Set population, find fitness of population, for individual apply selection, apply crossover and mutation, figure out and recreate chromosome. Approximately an optimized solution for large number of time will be provided.

Amitabh et al [16] invented a prioritization technique on the basis of binary code. They delivered a system called ECHELON. On the basis of modifications that are being done in the program it prioritizes the test cases. ECHELON is a unified part of Microsoft development process. It uses simple and quick algorithm. It give results within few seconds by saving time and resources.

H.Do et al [17] present a controlled experiment. Software developers used Junit framework for generating test cases that are being executed in java. Junit provides helps to testers to create test cases and to re execute these test cases when modifications occur in program. There experiment is for finding the efficiency and effectiveness of test case prioritization under this JUnit Framework. They developed 6 blocks and method level techniques which are as follows. 1) Total block coverage 2) Additional block coverage 3) Total method coverage 4) Additional method coverage 5) Total DIFF method 6) Additional DIFF method.

Bryce et al [18] presents the prioritization of test cases for interaction coverage. Their focus was for event driven software. On the basis of five criteria’s they prioritize the test cases. 1) Unique event coverage > Prioritize test cases like as soon as possible they measurer all different and uncommon events. 2) Event interaction coverage > Covers 2 way interaction and 3 way interaction. 3) Random test ordering > randomly ordering of test cases without any rule. 4) Shortest to longest with length of test cases. 5) Longest to shortest with length of test cases. The results concluded that for quick detection of faults, test suite must have dominant 2-way and 3-way interaction’s percentage.

Do et al [19] invented a technique in which they wanted to figure out what are the impact on specific prioritization technique of variations in time constraint and also the impact on cost of regression testing. They presented four techniques, in which two belong to total and additional coverage and two are related to Bayesian network. The equation used in this technique is:

$$COST=PS * \sum\nolimits^{n}_{i=2}(CS\ (i)\ +CO_i\ (i)\ +CO\ (i)+b(i)\ *\ CV_i(i)+C(i)\ *\ CF(i))$$

Additional techniques are considered to be better than total. Results have shown that time constraints perform a remarkable role in test case prioritization techniques.

Jieng et al [20] invented test case prioritization technique ART (Adaptive random). They presented nine new coverage based ART techniques. They categorized them into three groups “maxmin, maxavg, and maxmin”. Their coverage is at statement level, branch level, and function level. “1) ART-st-maximum 2) ART-st-maxavg 3) ART-st-maxmax 4) ART-fn-maximum 5) ART-fn-avg 6) ART-fn-maxmax 7) ART-br-maxmin 8)ART-br-maxavg 9)ART-br-maxmax.” A comparison was done between these techniques and randomly ordering and the results was these are 40-50% more efficacious than randomly ordering of test cases. ART-br-maxmax is perfect among all groups. In order of exposing defects and failures they are more effectivethan traditional coverage techniques.

Maia et al [21] invented a metaheuristic algorithm that is known as GRASP (greedy randomized adaptive search procedure). They have done automatic prioritization of test cases with the help of GRASP. A metaheuristic algorithm found best and optimistic solutions. They compared the GRASP technique with some search algorithms like simulated annealing, greedy, genetic. Their comparison was on the basis of performances and coverage. Their coverage criteria were block, decision and statement. The results has shown that additional greedy is best algorithm but GRASP is not worse than that among all these five algorithms. Among all these algorithms, GRASP surpassed the simulated annealing, genetic and greedy algorithm.

Dennis et al [22] invented prioritization technique using relevant slice. A program includes a lot of statements. Some statements have no impact on output generated by test cases but some statements have potential to effect the output generated the test. All these statements create a group and this group goes to relevant slice. In this technique following factors are considered. 1) No of statements of the output in relevant slice 2) No of statements of the output that are not in the relevant slice are implemented by test cases. Equation that is used for checking test case weight is,

$$TW = Reqslice + ReqExercise$$

Reqslice presents the no of requirements in the relevant slice of output. ReqExercise presents the no of requirements that are exercised by the test case.

Leung et al [23] invent a cost model that compares the certain regression techniques. They divide the cost into 2 groups. Direct cost and Indirect cost. Direct cost involves 1) System analysis cost Ca 2) Test selection Cost Cs 3) Test execution cost Ce 4) Result analysis cost CT. Indirect cost includes 1) Overhead cost 2) Tool Development cost. One Big disadvantage of this technique was that this technique ignores the cost of undetected faults.

For regression testing Alexay et al [24] invented cost model of cost benefits tradeoffs. They performed experiments for selection, reduction and prioritization of test cases. They used cost factors like

- Ca (T) analysis cost
- Ce (T) execution cost
- Cc (T) result checking cost
- Cs (T) selection cost
- Cm (T) maintenance of the test suite's cost.

In experiment for test case prioritization, they focus 2 factors which are cost required for analysis Ca (T) and cost of prioritization Cp (T). When they were performing experiments they divided testing process in two phase, one is preliminary and second is critical phase. These two phases are having different costs. The results have shown, optimal ordering, additional function coverage and total function coverage have maximum savings.

VII. Factor Based Comparison

On the basis of review of different prioritization techniques, in this a comprehensive table is developed. That is, in Table I, different papers are compared on the basis of factors.

TABLE I. FACTOR BASED COMPARISON OF TEST CASE PRIORITIZATION

S #	Methods/ Technique	Based on Factor	Key Point	Formula/Equation/Algorithm/Tool/Technique Used
1.	Hema Srikanth and Laurie Williams	Based on Customer Requirement	Test cases are being ordered according to WP values. WP is weightage Prioritization. Test cases having the higher values are executed first.	$WP = €(PF\ value * PF\ weight)$
2.	R.kavitha et al.	Based on Customer Requirement	they consider 4 factors 1) Priority of requirements assigned by customer 2) Code implementation complexity assigned by developer 3) Changes in requirements 4) Fault impact Test cases are ordered according to values of TCW. TCW represents test case weight.	$RFVi = \sum_{j=1}^{3} Factor\, value\ j / 3$ $TCW = \left[\frac{\sum_{x=1}^{i} RFVx}{\sum_{y=1}^{n} RFVy} \right] * i / n$
3.	Ashraf et al.	Based on Customer Requirement	they consider 6 factors 1) Modification in requirement 2) Priority of customer 3) Complexity in Implementation 4) Traceability in requirement 5) Time of execution 6) Impact of fault	They present a value based prioritization algorithm. To get the net values calculations are being done on the values get from the above 2 levels. These values are further used for ordering of test cases.
4.	Wong et al.	Based on Code Coverage	Propose a technique in which their criterion of test case prioritization is of increasing cost per additional coverage.	They use the tool called ATAC an automatic testing tool for analysis in c.
5.	Rothermal et al.	Based on Code Coverage	Propose 4 coverage based techniques they are total coverage, additional, branch and	they used the Aristotle a program analysis tool. APFD is used

			statement coverage.	for measuring the results.
6.	Erlbaum et al.	Based on Code Coverage	Propose the version specific prioritization technique. They present 8 function level techniques they are 1) Total function 2) Additional function 3) Total FEP function 4) Additional FEP function 5) Total FI function 6) Additional FI function 7) Total FEP FI functional 8) Additional FEO FI functional	APFD metric is used for fault detection. ANOVA and Bonferroni analyses were performed on all techniques.
7.	Amitabh Srivastava and Jay thigarajan	Based on Code Coverage	Proposed a prioritization technique based on binary code.	They gave a system called ECHELON. □ ECHELON prioritize the test cases based on the modification are being done in the program.
8.	Belli et al.	Based on Code Coverage	Proposed techniques in this ordering of relevant events are being done. The events have many features. Events are prioritizing according to the importance of their features.	Graph modal based approach is used for prioritization. Fuzzy c-Mean clustering algorithm is used for erection of events.
9.	Do et al.	Based on Code Coverage	Proposed a technique for regression tetsing in which they want to find out what are the effect on a specific prioritization technique of variation in time constraint and also the effect on cost profit.	$COST = PS * \sum_{i=2}^{n} \left(CS(i) + CO_i(i) + CO_r(i) + b(i) * CV_i(i) + C(i) * CF(i) \right)$
10.	Leung and white	Based on Cost	Propose a cost modal that compare the various regression strategies. They divide the total cost into two parts • Direct cost • Indirect cost	Direct cost includes 1) System analysis cost Ca 2) Test selection cost Cs 3) Test execution cost Ce 4) Result analysis cost Cr Indirect cost includes 1) Overhead cost 2) Tool development cost
11.	Alexey Malishevsky et al.	Based on Cost	Proposed cost modal of cost benefit tradeoffs. They did experiments for selection, reduction and prioritization and presents cost modals for them.	In experiment for test case prioritization they consider two factors cost required for analysis and prioritization $Cp\ (T)$. They divide the testing process in two phase. These two phases have different costs.
12.	Jung-Min-Kim and Adam Porter	Based on chronographic history	It is for regression testing. Their main motive behind this is to show that historical information can be useful for decreasing the cost and it may be beneficial in increasing the efficiency of testing process.	They did comparison of some prioritization methods like LRU, random, safe random. Weakness of their cost modal is that they only take the consequence of last execution of the test case
13.	Fazlalizadeh et al.	Based on chronographic history	They make some changes in the technique of Kim and porter. If resource and time constraint environment is considered they motive is to give faster fault detection.	A comparison was being done with the random ordering. The box plots shows that it has faster fault detection and stability.
14.	Park et al.	Based on chronographic history	Propose an approach for cost-cognizant test case prioritization. that uses the historical information.	A comparison is being done between their technique and functional coverage technique. Results show that in terms of APFD it better than functional level technique.

VIII. Comparison of Different Techniques

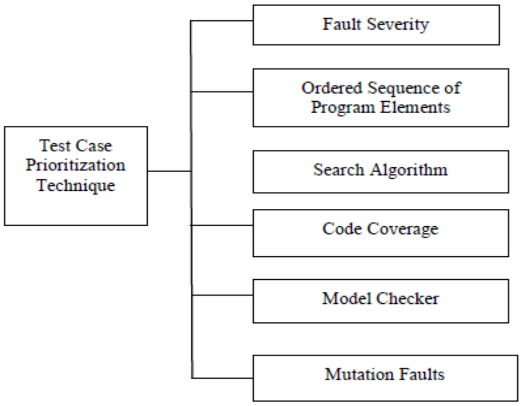


Fig. 6. Techniques of Test Case Prioritization

Figure 6 is showing different techniques of test case prioritization. As, in software development life cycle regression testing is very expensive process. But it makes ensure that the project will satisfy all requirements of stakeholders. In testing phase, about 50% of the total software cost is consumed [25]. Engineers perform assigning test cases preferences through regression testing and execute those test cases which have more significance. The main target of test case prioritization is fault detection. In software testing now a days there are so many techniques which are invented by different researchers to prioritize the test cases. Different techniques are compared which are widely used by the researchers these days. Such techniques are mutation faults, model checker, ordered sequence of program elements, fault localization, fault severity etc. Each technique has own pros and cons. In this section advantages, disadvantages and main idea of techniques are discussed, and on basis of that graph based results is generated.

TABLE II. Comparison of Test Case Prioritization Techniques

S.No	Technique	Key Idea	Advantage	Disadvantage
1.	Fault Severity	Base on Requirement specification	1. It enhances the software quality. 2. The faults are discovered quickly with high severity. 3. It can enhance the fault detection rate. 4. Requirements volatility is most important factor.	1. It does not remove the induced factor of requirements volatility. 2. Project scope is limited
2.	Fault Localization	Based on execution information of fault localization.	1. A postmortem analysis approach. 2. Faster failure exposes	1. It is not much effective. 2. The subsequent fault localization may suffer
3.	Mutation faults	Based on changes in program code	1. Fault detection rate is Improved	1. Cost reduction is still not significant.
4.	Ordered Sequence of Program Elements	Based on execution frequencies of the program element	1. Bugs are detected quickly in loops. 2. Cost effective approach	1. Still it's not much effective approach.
5.	APFD	Based on average faults found per minute	1. Rate of faults detection is easy at system level	1. This technique not much more efficient in fault detection.
6.	Model Checker	Based on functional model of program test	Prioritization is efficiently applied on the time of creation of test cases	Many factors are still not included such as: 1. Actual test case execution costs. 2. The costs of potential Faults
7.	Search Algorithm	Based on code coverage	1. Efficient 2. Flexible. 3. Program's size does not have impact on prioritizing the test cases.	1. Still cannot solve large number of test case. 2. Sometimes it produces different results.

IX. Findings and Results

There are many other techniques that are used for test case prioritization such as Empirical study, coverage based, Decision coverage etc., but did not discussed in this paper. Figure 7 is showing the graph result that is the comparison of seven techniques which are commonly used now a day. Each and every technique has its advantages and disadvantages. These techniques are based on different factors. If the tester wants to pick any technique so he/she can choose any technique according to his/her requirements and specifications.

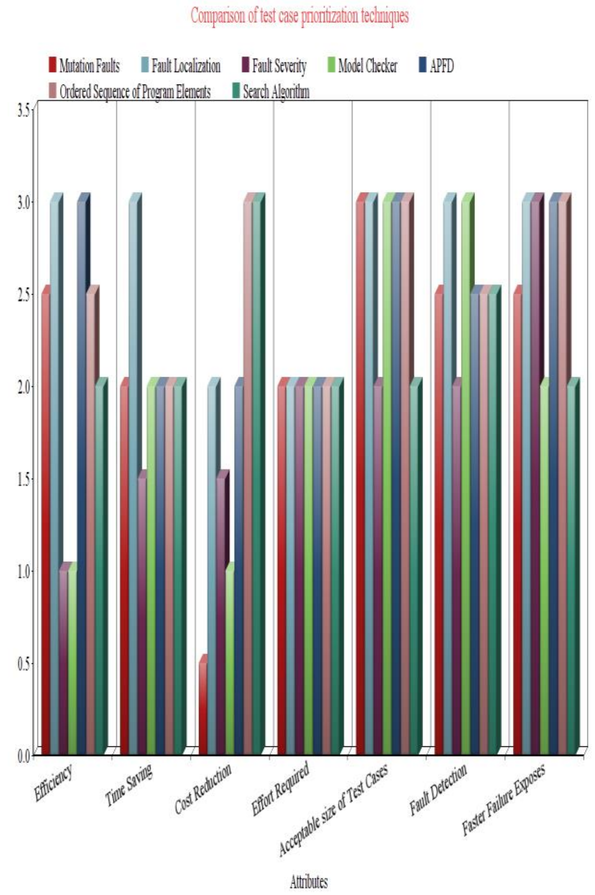


Fig. 7. Comparison of Test case Prioritization Techniques

Testing is a technique for evaluating product quality and also for indirectly improving it, by identifying defects and problems. The testing is conducted in view of a specific purpose (test objective), which is stated more or less explicitly, and with varying degrees of precision. Stating the objective in precise, quantitative terms allows for establishing control over the test process.

X. Conclusion

Testing is an important and mandatory part of the software development. Software testing is most significant process of software development life cycle. The testing phase involves finding the bugs and removal of defects at earliest if possible. There are different types of testing that software tester adopt according to their requirements such as Mutation, Regression, Stress, Security, Load testing etc. Regression testing is the important type of software testing. When modifications occur in software then there is a need to perform regression testing to check that it doesn't influence the other modules of system. Test case prioritization is done by using different techniques. This paper furnished a comprehensive analysis of different various regression techniques, which primarily focuses on prioritization of test cases. Prioritization means "To schedule or organize" the test cases execution. Few Prioritization techniques are examined in details. Software quality and fault detection in order to more effectiveness and efficiency can be enhanced through regression testing. This paper comprehensively summarizes different research articles (via practitioners) along with their techniques, approaches and methodology they used. Many techniques are investigated and compared that are used for test case prioritization. Each technique has its own advantages and disadvantages. All techniques are tried to explained and concluded so that tester can use any technique according to their requirements and need.

References

[1] S.Yoo, M.Harman, "Regression testing Minimisation, Selection and Prioritization: A Survey" Wiley InterScience DOI: 10.1002/000,2007

[2] Sahil Gupta, Himanshi Rapria, Eshan Kapur, Harshpreet Singh, And Aseen kumar " A Novel Approach for Test Case Prioritization " at IJCSEA, Vol. 2, No.3, June 2012

[3] Thillaikarasi Muthusamy and Dr. Seetharaman.K "EFFECTIVENESS OF TEST CASE PRIORITIZATION TECHNIQUES BASED ON REGRESSION TESTING" at (IJSEA), Vol.5, No.6, November 2014

[4] Hema Srikanth, Laurie Williams, "Requirements-Based Test Case Prioritization".

[5] Siripong Roongruangsuwan and Jirapun Daengdej. Test Case Prioritization Techniques. JATIT, 2005-2010.

[6] Praveen Ranjan Srivastava. Test Case Prioritization. JATIT, 2008.

[7] B. Korel, G. Koutsogiannakis, "Experimental Comparsion of Code Based and Model model Based Test prioritization," IEEE 2009.

[8] B. Korel, G. Koutsogiannakis, and L.H.Tahat, "Application of System Models in Regression Test Suite Prioritization," in Proceedings of the 24thIEEE International Conference Software Maintenance (ICSM '08) pp.247- 256, 2008.

[9] E.Ashraf, A.Rauf and K.Mahmoat "Value based regression test case prioritization"WCECS 2012, October 24-26 2012.

[10] kumar, Dr Varun,"Sujata and M.kumar,"Test case prioritization using fault severity""IJCST 1, no.1 (2010):67-71.

[11] R.Beena, S.Sarala "code coverage test case selection and prioritization" IJSEA vol.4, no.6, November 2013.

[12] Prakash.N, Rangaswamy "Potentially weighted method for test case prioritization" JCIS 9:18(2013) 7147-715

[13] Praveen Ranjan Srivastava "Test case prioritization", Journal of theoretical and applied information technology, 2005-2008.

[14] R.Kavitha, N.Sureshkumar "Test case prioritization for regression testing based on severity of fault", IJCSE vol.02, no.05 2010, 1462-1466.

[15] Ruchika malhotar, abhishek bharadwaj "Test case prioritization using genetic algorithm" IJCSI: 2231-5292, vol-2, Issue-3, 2012.

[16] A.Srivastava, and J.Thiagarajan, "Effectively prioritizing tests in development environment", Proceedings of the International Symposium on Software Testing and Analysis, pp.97-106, July 2002.

[17] H.Do, G.Rothermel and Kinner, "Empirical studies of test case prioritization in a Junit testing environment", Proceeding of the International Symposium on software Reliability Engineering, pp.113-114, NOV 2004.

[18] R.C. Bryce, A.M. Menon, "Test Suite Prioritization by Interaction coverage", Proceedings of the workshop on domain specific approaches to software test automation (DOSTA), ACM, pp. 1-7, 2007.

[19] H. Do, S. Mirarab, L. Tahvildari, G. Rothermel, "An Empirical Study of the effect of time constraints on the cost benefits of regression testing" Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering,pp71-82,2008.

[20] B.Jiang, Z.Zhang, W.K.Chan, T.H.Tse, "Adaptive Random test case prioritization" In Proceedings of International Conference on Automated Software Engineering, pp:233-243, 2009.

[21] C. L. B. Maia, R. A. F. do Carmo, F. G. De Freitas, G. A. L. de campos and J. T. De Souza, "Automated test case prioritization with reactive GRASP," In Proceedings of Advances in Softwar Engineering, pp.1-18, 2010.

[22] Dennis Jeffrey and Neelam Gupta “Test case prioritization using relevant slices” Department of computer science The University of Arizona TUCSON, AZ85721

[23] Harton K. N. Leung and Lee White “A Cost Modal to compare Regression Test Strategies” CH3047-8|91|0000|0201, IEEE 1991.

[24] Alexey G. Malishevsky, Joseph R. Ruthruff, Gregg Rothermel, Sebastian Elbaum “Cost-cognizant Test Case Prioritization” Technical Report TRUNL-CSE-2006-0004, Department of Computer Science and Engineering, University of Nebraska – Lincoln, 2006.

[25] Srikanth, Hema, Laurie Williams, and Jason Osborne. "System test case prioritization of new and regression test cases." In Empirical Software Engineering, 2005. 2005 International Symposium on, pp. 10-pp. IEEE, 2005.

The Impact of Analytical Assessment of Requirements Prioritization Models: An Empirical Study

Aneesa Rida Asghar
Dept. of software engineering
Bahria University Islamabad, Pakistan

Atika Tabassum
Dept. of software engineering
Bahria University Islamabad, Pakistan

Dr. Shahid Nazir Bhatti
Department of Software Engineering
Bahria University Islamabad, Pakistan

Dr. S Asim Ali Shah
Dept. of Electrical Engineering
Bahria University Islamabad, Pakistan

***Abstract*—Requirements prioritization is one of the important parts of managing requirements in software development process which plays its role in the success or failure of a software product. A software product can go wrong or fail if right requirements are not prioritized at right time. Thus, there is a need of a vast or complete requirements prioritization technique or model that spans all the factors that must be considered while prioritizing requirements whether it's for a traditional software development or agile software development. There are several requirements prioritization methodologies that aid in decision making and in prioritizing requirements but importantly many lacks to account the important factors that have significant influence in prioritizing requirements. A requirement prioritization methodology that takes account of important factors such as time and human behavioral factors that has an influence in prioritizing requirements is required. This new model/ technique expected to overcome the lack that is in existing prioritization techniques because of not considering time gap factor and human behavioral factor. Extensive study on literature of agile methodology, requirements elicitation and prioritization has been done to find out factors that influence the decision making process of requirement prioritization. It is found that as agile methodologies such as XP, SCRUM deliver products in increments, there is a time gap between each increment of approximate 4 weeks or more, this time lapse could cause human behavioral to change either because of market demand or any other personal reason and, thus, influences the prioritization decision. These factors could be termed as time factor and human behavioral factors. Thus, a requirement prioritization technique or model is needed that takes account of all such factors while prioritizing requirements whether it's for a traditional software development or agile software development.**

Keywords*—Agile Software Engineering (ASE); Agile Software Development (ASD); Scrum Software Development Process; SCRUM; Product Owner (PO); Extreme Programming (XP); Requirements Prioritization techniques; Analytical Hierarchy Process (AHP); Cummulative Voting (CV); Numerical Assignment (NAT)

I. Introduction

Managing requirements is one of most important aspect of software development system. Developing software is entirely based on requirements as it contains the functionality or quality that the customer or stakeholder/s needs. Requirements emerge throughout the development process of software and, thus, they are needed to be addressed properly through communication between stakeholders, developers and documentation. A lot of factors play their role in the success or failure of a software product such as eliciting right and unambiguous requirements, managing unrealistic requirements and focusing on quality requirements etc. Requirements prioritization is one of the important parts of managing requirements in software development process which plays its role in the success or failure of a software product. A software product can go wrong or fail if right requirements are not prioritized at right time. Thus, there is a need of a vast or complete requirements prioritization technique or model that spans all the factors that must be considered while prioritizing requirements whether it's for a traditional software development or agile software development. Agile methodology is an innovative and iterative process that is currently the most widely used methodology for software development around the world as it supports changing requirements and helps in addressing changes throughout the development process.

There are many existing requirements prioritization methodologies that aid in decision making and in prioritizing requirements but importantly many lacks to account the important factors that have significant influence in prioritizing requirements. A requirement prioritization methodology that takes account of important factors such as time and human behavioral factors that has an influence in prioritizing requirements is required. This new model/technique is expected to overcome the lack that is in existing prioritization techniques because of not considering time gap factor and human behavioral factor. Agile methodology such as XP, SCRUM delivers software products in increments; especially in case of SCRUM, since there are time gaps between sprints, human behavioral factor plays an important role here as the time passes and the requirements changes. In sprints, requirements will be prioritized both on the basis of influencing factors such as cost, value, risk, time to market etc. and through the effect of non-functional requirements over functional requirements. This will improve the overall quality of software product when it is included in the development process of scrum or could at least reduce the wastage of time, effort and resources. Requirements will not only be prioritized based on sprints, human decision but by critically analyzing the factors (sub characteristics) that can cause the product to success/ fail repeatedly thus ensuring the consistency in right requirements and hence the right prioritized requirements will be selected for a particular sprint at a time.

Problems arise when new requirements evolve due to change in business needs, time to market, time and human behavioral factors during the development of a software product [4] [3] [1]. This is why the software market is moving towards an approach that supports changing requirements and managing them. As agile software development contains this attribute of managing changing requirements it is being widely used in software developments process worldwide where speedy development process is required. There are many factors involved in the success or failure of a product, one of them is collecting and prioritizing requirements while keeping influencing factors in mind [2]. After carefully eliciting requirements it is essential to arrange them so right requirements are delivered at right time in order to ensure the success of a product. There are many well-known existing requirements prioritization techniques but not one of them spans all the different types of software development projects. Some techniques work well with short projects and some with large projects, some with traditional development where extensive documentation is needed with no changes during the development process and some with agile development where little or no documentation is needed and where changes are welcomed. Although requirements can evolve at any stage during the development process, it is very unlikely to be able to handle the factors that could be the cause of new emerging requirements or the cause of changing in existing requirements. But what could be done is that a new method or techniques could be introduced that considers those factors which are expected to be the root cause of these changes which lead to the waste time and effort; and thus reduces the wastage of time, effort and resources or could at least minimize the damage.

Requirements are elicited at the beginning of every software development process and project (product) and later are prioritized according to their relative importance to the market and to the product itself by keeping several factors in mind that could affect the prioritization decision. Prioritizing right requirements at right time helps the software team to understand the existence and importance of a particular requirement, its importance of use and its urgency to time to market and many other factors. There are many existing requirements prioritization techniques with their relative strength and weaknesses depending on many aspects they consider while prioritizing requirements. However, many of them fail to take account all the factors that must be considered while prioritizing requirements such as cost, value, risk, time to market, number of requirements and effect of non-functional requirements on functional requirements, time constraints and human behavior factor.

One of the most popular methods among agile family where software is delivered in increments called sprints is known as SCRUM [8] [6]. A sprint consists of 2-4 week iteration. Scrum methodology comprises of a planning meeting and daily scrum meeting, the planning meeting is conducted at the beginning of every sprint. In this meeting team members determine the number of requirements they can oblige to manage that is they create a sprint backlog out of that. Sprint backlog contains the list of all the tasks that should be perform during a particular sprint. Daily scrum meetings are not more than 15 minutes, where product owner (PO) gets continuous updates about the development process and can provide feedback about the features being included. This way if a PO decides to add new feature to a sprint, he/she can discuss it with the development team and save time rather than reviewing it at the end and demanding change at the end. The team conducts a sprint review at the end of each sprint where they demonstrate new features and functionality to the PO or to other stakeholders that can provide any kind of feedback which could be beneficial or helpful in any way for the next sprint. This loop of feedbacks results in modifications to the recently delivered functionality, then again it is more likely reviewing or adding new requirements to the product backlog. Another activity in Scrum project management is Sprint retrospective. The Scrum Master, PO and the development team participates in this meeting. It is the chance to reproduce or review the sprint that has ended, and identify new ways to improve. Scrum consists of three artifacts, sprint backlogs, product backlogs and burn down charts. The Product backlog, prioritized by the PO is a complete list of the functionality (written as user stories) that is to be added to the product eventually. It is prioritized so that the team can always work on the most important, urgent and valuable features first. On the other hand, sprint backlog is the list of all those tasks that the team obliged to and needs to perform during the sprint in order to deliver the required functionality. The remaining amount of work either in a sprint or a release is shown by ‘Burn down’ charts. It is an effective tool to conclude whether a sprint or release is on schedule to have all planned work finished in time. The traditional requirements engineering is very time consuming and requires speedy process to timely meet the needs of market so modern software industry demands rapid and iterative process like agile development to cope with the changing requirements and time.

As XP, SCRUM and other agile methodologies allow engineers to handle changing requirements as they evolve; however, it is still a challenging task to comprehend which prerequisites are sufficiently vital to have high need and to be incorporated into early sprints because later on this decision could be influenced by other factors which particularly in case of SCRUM could be time gap and human behavioral factors. Organizing requirements into Priority requirements helps the project team to comprehend which requirements are most essential and most urgent to implement and execute. Prioritization is likewise a helpful activity for decision making in other phases of software engineering. Therefore there should be a well-managed requirement prioritization technique included in scrum processes that minimizes changes later in the process and save time, effort and other resources.

II. Literature Review

In this work [1], author presented the 10 years progress of agile research and proposed some future research areas for agile researchers to hold on to an approach that is theoretical or hypothetical. A survey based methodology was used to get reliable information about the progress of agile methodologies. It is significant to remember that one can produce and enhance fields as a scientific discipline only if energies are able to convey a solid theoretic system to conduct research on agile development. Therefore, it is a need that in future when investigating into agile development proficient research areas, agile researchers hold on to a more theoretical based approach.

Ming Huo et al [3] proposed that agile methods can assure quality even agile methods are faster and have to manage changing requirements. Author basically presented a comparison between waterfall model and agile model and presented the results. Agile methods contain some practices that have QA abilities, so with the help of this quality can be achieved more appropriately through agile methods. However one thing that must be considered when documenting agile RE is that in complex software development processes, less documentation can bring some issues/ problems.

Lan Cao et al [4], presented an empirical study on agile RE practices. This study shows the difference between agile RE and traditional RE is an iterative finding approach. Developing clear and complete requirements specification is impossible in agile development. Because of such important differences a new set of agile RE practices had come into practices that are reported in this paper. The study participants recognized that the most important practice in RE is thorough communication between the developers and customers.

Numerous participants highlighted that the efficiency of this practice depends deeply & effectively on exhaustive communication and interaction between customers and developers. Risks such as incomplete requirements, ineffectively developed requirements or wrong requirements are possessed if high quality interaction lacks in any project.

In this work Pekka et al [6], proposed that there are different methods of agile process that needs the empirical evidences. Authors emphasized on the quality of methodology not the quantity. This approach was chosen for comparative analysis of these processes. Five perspectives are included in the analytical lenses. SDLC include the process aspect abstract principles vs. concrete guidance, empirical evidence, project management and universally predefined vs. situation appropriate. New directions are offered based on these 5 perspectives that focus on quality not on quantity of methods.

Amin et al [7], proposed that some lessons of RE must be considered by the agile methods if the most emphasized thing is quality. Some major aspects of RE that are not a much emphasized in agile are analysis (verification and validation), non-functional requirements and managing change. Author suggested that these practices of RE can be adopted in agile and high quality can be achieved. RE practices such as simplicity, continuing validation, short releases and frequent refactoring, can be implemented in the perspective of agile main ideas.

Deepti Mishra et al [8], proposed that agile process can be helpful for the development of complex software projects. Author supported his argument with the help of a case study. A medium enterprise (SME) that practiced agile methods, achieved many successful results. Starting a project with agile methods and then achieving optimum methods by tailoring agile methods according to vision and benefits is the main reason of the success of supply chain management. The architectural design of this large scale complex project was supported with formal documentation. In the successful completion of the project an important role was played by this design documentation.

Franek et al [11], proposed different ways of RE methods from which agile software development can get advantages. Some common and different features and attributes of traditional approaches and agile approaches are also discussed. Agile approaches such as XP involves feedback from development teams and customers, communication and simplicity. Similarly RE process also includes dictation, analysis and validation. But in agile process phases are not as clearly distinguished as in RE process and techniques can also vary. Overall both are pursing same objectives. The major difference is of documentation that is really important to communicate with the stakeholders.

V. N. Vithana [12], conducted a research using qualitative methods to find out which requirement engineering practices are mostly being used in SCRUM methodology when developing a software product offshore. In order to collect data different job holders from nine organizations were questioned. It was found that RE practices such as Customer Involvement, prototyping, test driven development and Interaction are the least practiced activities of Requirement Engineering, although most of the team members were successfully practicing iterative requirements engineering, face to face communication, managing requirements change and requirements prioritization of SCRUM RE practice.

In this Anna Perini et al [14], proposed a strategy called Case-Based Ranking (CBRank). This method joins the preferences of the stakeholders of the project with the approximation of requirements ordering that is computed over machine learning methods. On simulated data the properties of CBRank are performed and then matched with a method called state-of-the-art prioritization, thus provided empirical results. However there are some assumptions in the CBRank prioritization process such as arbitrary selection as pair sampling policy and the monotonicity of the elicitation process. To improve the efficiency of real complex sitting methods the authors intend to work in future on non-monotonic formal logic case and pair sampling strategies that are more refined.

DAN HAO et al, [10] in this article, have presented a strategy that comprises the total and additional strategies for unified test case prioritization. These tactics prioritize test cases in light of components secured per test case, the aggregate number of program segments (or code-related) and the number of others (not yet covered) program segments (or code-related) components covered per test case, respectively. The proposed approach includes basic and extended models, which define a spectrum of test case prioritization from a purely total to a purely additional technique by specifying the value of a parameter referred to as the fp value [10].

Rahul Thakurta [15], proposed a quantitative structure that determines the priority of a list of non-functional requirements. This framework involves members from business organization and the project to provide a measurable ground for assessing the level of value addition that is considered while choosing a new non-functional requirement to the project's requirement set. However, the inputs provided to the framework by members were subjective which may result in non-optimal results. Additionally, as the requirements assessment process involves stakeholders from both business organizations and the project, there are odds of irreconcilable interests and priorities of requirements. The author has also set the directions for future work which is to build a heuristic to bind the number of stakeholders to be preferred for assessment process.

Naila Sharif et al [16], devised a new requirements prioritization technique called FuzzyHCV which is a hybrid of two domains (SE and Computational Intelligence). It is a fusion of two methodologies which are Hierarchical Cumulative Voting (HCV) and Fuzzy Expert System. In FuzzyHCV, rather than a single crisp value a triangular fuzzy number is used. The proposed technique has been applied on 3 case studies and the results obtained are very close to the results of actual prioritization used in all of the three case studies. It is found that FuzzyHCV produces more precise results than HCV by comparing them with actual results for the chosen datasets. Authors intend to carry on work in this area by using fuzzyHCV for other domains problem such as decision making problems in employee selection and by incorporating fuzzyHCV to already existing decision making or requirements prioritization techniques so that less risky choices are made in future.

Mukhtar A. Abo Elsood et al, 2014 [10] in this research paper conducted a survey on the most popular requirements prioritization techniques being used and their reported drawbacks. The authors have devised a goal-based requirements prioritization technique that is based on generating a relative weight for the requirements with respect to the identified goals by stakeholders after conducting the survey. This technique is expected to overcome requirements prioritization problems such as time consumption, scalability and complexity. This technique has been evaluated by a case study and has been compared with AHP; it has proven to be more effective than AHP. However this technique has only been compared with AHP and not others, which leaves the effectiveness of this technique into an unanswered question. The authors intend to solve problems of data vagueness and uncertainty by enhancing goal-based RP technique.

Mr. Seyed Ali Marjaie et al [11] stated in this paper that there are many factors in the requirements prioritization process which have not been observed carefully other than risk, cost and value; these factors have significant impact on prioritization result itself. A statistical method has been proposed by authors which is based on attributes such as elicitation, numeral assignment, and factor analysis. This method combines two or more attributes into a single factor thus reducing the number of attributes and tries to identify groups of inter-related attributes, to find out how they are related. This improves the stability of factors involved in prioritization process and also the existing prioritization techniques effectively. Attributes that have been selected as important attributes are cost, time, risk, reuse of code, complexity, desirability and frequency. However this proposed method has not yet been applied on any real time software project and the results are merely based on theoretical assumptions.

Nikita Garg et al, 2015 [12] in this research paper explained all the requirements elicitation and requirements prioritization techniques. The requirements prioritization techniques that have been discussed in this paper are Analytical Hierarchy Process, The 100 Dollar Test, Numerical Grouping, Ranking and Top-Ten requirements. The authors have explained why it is important to select right requirements elicitation and prioritization techniques when building software as it acts as a backbone for the project. The authors have explained how each type is suitable for a particular situation but have not compared any two techniques nor they have suggested any new or hybrid technique.

Muhammad Imran Babar et al [13] to overcome the limitations of existing software requirements prioritization techniques, the authors have proposed an extension in VIRP model. The proposed technique will be automated for better understanding by adding heuristics using Neural Network and thus the interpretation of important requirements will be better so that there would be less chances of error. It is expected to be more time efficient, scalable as well as high overall performance than other techniques. However the proposed technique has not yet been implemented and the results are just expected to be good when compared to other techniques, there is no validity of this proposed technique.

Richard Berntsson Svensson et al, 2011 [14] in this paper found out that the dominant method that is being used in different companies developing software intensive systems are ad-hoc prioritization and priority grouping of requirements. The authors conducted a survey in 11 successful software companies. They also found out that customer input was being used as criteria for prioritization but not all the time. The results also suggested that functional requirements were given more importance than the quality requirements. The non-functional requirements (quality requirements) are prioritized if only time and resources are still available after implementing functional Requirements.

M. Waseem Asghar et al, 2013 [15] in this paper devised a tool called SWTMetrics. Using artifacts traceability information this method prioritizes changing requirements. A set of code-based metrics is also being used to locate requirements implementation as well as it measures several properties of requirements being implemented such as size, coupling, scattering. Authors have applied the proposed tool on three java applications and the results achieved are considerably different than those defined by experts but not entirely. This diversity is because of analyst selecting those requirements that are weakly related to main functionality (provided by the application) with respect to SWTMetrics. Hence, the tool determines the ordering of requirements based on how these are implemented in a subject software system but its effectiveness has not yet been confirmed by the importance of being applicable in the software industry and providing some promising improved results.

Muhammad Aasem et al [8] proposed a framework in this research paper that combines existing approaches and techniques to help software engineer in performing prioritization. The proposed framework has α, β, and γ processes where the first two processes α, β, are subjective and require human involvement; they (α, β) include 100-dolors test prioritization method. Whereas, because of the algorithmic nature of the third process γ; it can be made fully automated as by using AHP technique it can automatically perform pair wise comparisons when the outputs produced by process-α and process-β are in the form of batches of n size each and Ranked Criteria respectively. Release scheduler sub-process of process-γ is executed (which is based on Numerical Assessment) after mapping all requirements into B-Tree. Thus a series of releases of prioritized requirements is obtained. This framework is expected to be effective but has not yet been tested on real scenarios. Feasibility of processes α and β for semi or full automation should also be checked.

Nupul Kukreja et al [17], in this have proposed a prioritization methodology to prioritize requirements of system and software. This methodology is a two-step approach and is based on decision theoretic model using a prioritization algorithm called TOPSIS viz. In the proposed approach [13], initially, the system is fragmented into high-level Minimal Marketable Features (MMFs). The proposed methodology allows measuring the effect of fluctuating business priorities on individual requirements without much overhead. This methodology also authorizes stakeholders to perform numerous analyses which also help in accurately judging the impact of fluctuating business priorities on individual requirements.

Here authors have also presented a validation report of this methodology by implemented this with 24 project teams of students at the Software Engineering project course in the University of Southern California. Although this approach has some drawbacks that need to be tackled in future; such as, one of the drawbacks of TOPSIS is reversal of ranks i.e. the original order of requirements prioritization may change if irrelevant requirements are entered into the prioritization. This limitation was not considered while implementing the approach as the teams were result oriented therefore they resisted in adding irrelevant requirements for prioritization. Another drawback is that the ordered prioritization of requirements may not accurately reflect the anticipated rank ordering of requirements

To overcome the drawbacks of TOPSIS, several other prioritization algorithms could be used instead of TOPSIS viz such as Cost of Delay, Simple Additive Weighting or Weigers' Prioritization. Also one can simply record items to eliminate the overhead of winbook's incapability to record the items.

III. Related Work

Missing or poorly specified quality requirements can lead to project failure or huge loss. Eliciting quality requirements effectively is a difficult task altogether especially in SCRUM where one person i.e. the product owner [PO] has to make the list of all the requirements to be included in the project. It can be a hectic and difficult task. As 'Quality' requirements drive the architecture of software-intensive systems, they are more important than the functional requirements. Thus the success or failure of mission critical systems depends on how well the quality requirements are engineered and implemented. Prioritizing requirements is also another challenging task while developing a software product. Product Owner's commonly use following backlog prioritization techniques: Kano analysis, Moscow and Relative weighting (Karl wieger) [3] [8].

A. Analytical Hierarchy Process

In AHP the priorities of requirements is calculated to estimate their relative importance by comparing all unique pairs of requirements. In other words, the individual performing the comparison will decide manually which requirement has more significant, and to what extent using a scale 1-9.[14] AHP provides better results than any other tested methods as it is a ratio scale methodology, and also includes a consistency check.

Steps involved in AHP are:

1) Make an v×v matrix (v represents the number of requirements) requirements are latter inserted in rows and columns of the matrix.

2) For each pair of requirements, insert their relative intensity of importance (where the row of X meets the column Y). At the same point, insert the reciprocal values to the transposed positions (e.g. if cell XY=4 then cell YX=1/4)

3) Now, calculate the eigenvalues of this matrix to get the relative priority of each requirement. The final result will be the relative priorities of the requirements.

Total no. of comparisons that AHP requires is $v×(v−1)/2$. Redundancy is produced in pair-wise comparisons in AHP, therefore AHP also calculates the consistency ratio to check the accuracy of the comparisons [14].

B. Cumulative Voting (CV)

CV is a ratio-scale requirements prioritization technique where the customers/stakeholders are given a fixed number of 'units' which are used for prioritization of requirements by giving vote to the requirements that the customers/stakeholders think are important or delivers the highest functionality. Another important feature of CV is the 'weightage'. For example there are 3 stakeholders then; Stakeholder with highest authority/share is given the highest weight (e.g 10) and the stakeholders 2 and 3 with lower shares are given lower weights e.g. 7 and 5 respectively. Their weight is multiplied by the number of units the stakeholders assigns to requirements. In this way if stakeholder 2 and 3 vote for a particular requirement say 'reqA' which stakeholder 1 does not vote for and instead votes for ‘reqB’ then ‘reqB’ will be of high priority even though reqA got 2 votes and reqB got only 1 vote.

C. MoSCoW

MoSCoW stands for Must have, Should have, Could have and Would have requirements. It is based on human opinion, based on their experience, desire and influencing factors at that time such as market demand, cost, risk, time and resources. Must have requirements are critical to the current increment in order to be a success, these are time critical requirements. Should have requirements are important to be included in the product but are not necessarily important to add to the current increment and can be added on later increments, these are not time critical requirements. Could have requirements are nice to have in the products but they do not participate in the success or failure of the product but are still nice to have if included. These requirements could improve user experience or satisfaction. Won’t have requirements are the ones that are least critical to time or success of the product and hence can be added later any time of the time and resources permit.

D. Numerical Assignment

In Numerical assignment numerous requirements are grouped into different priority groups such as high, medium and low priority groups. All the requirements in a particular group will have same priority. For example if there are 7 requirements in medium category then all these 7 requirements will have same priority for this group.

E. Bubble Sort

In bubble sort prioritization, two requirements are taken and then compared manually; if the person doing the comparison feels that 1st requirement should have higher priority than the other requirement then he/she swaps the priority and continues this process until all the requirements have been compared. The result will be a prioritized set of requirements.

F. Ranking

In simple ranking, requirements are ranked manually from 1 to n number. 1 being the highest priority rank and n (the last integral valued requirement) being the lowest priority rank.

G. Hundred Dollar Method

Hundred dollar method is sometimes considered as Cumulative voting, however it is different than CV in a sense that ‘weight’ is not assigned to stakeholders in hundred dollar test. Each stakeholder is expected to distribute 100 dollars to the set of requirements being considered; however one may distribute full 100 dollars to a single requirement if he/she feels it’s the most important requirement but is being neglected.

H. Binary Search Tree

In binary search tree, each node represents a requirement. Each base node has two child nodes with lower priority requirements on left child node and higher on right child node. We take a random requirement and compare it to the root node. If that requirement has lower priority than the root node then it is compared to the left child node; and now if it has higher priority than this child node it is placed on the right side of this node, however if it has lower priority then it is placed to the left side or compared to the left side child; or placed to the left side of the root node of there is not already any child on the left side of root node, otherwise if it has higher priority than it is compared/placed on right side of the root node and so on. This process of comparing nodes (requirements) to the root node and so on is done until all the nodes have been placed in their right priority.

I. Five Whys

Often stakeholders want a certain requirement implemented which does not have any great function or quality aspect and does not even have founded on logical arguments or the business interests of the company but still keep on insisting on that particular requirement. In such case, the team members (engineers) ask 5 whys (repeatedly 5 times or less) to why this requirement is important enough for the stakeholder to be implemented until the importance of the requirement is either found or established. The answer found could either determine the priority/importance of the requirement or that it could be cancelled or postponed for later increments.

IV. Proposed Methodology

The proposed model is based on several techniques that are being used to prioritize requirements. However when combined, they are expected to give better results. The First step in this model is cumulative voting, in cumulative voting each stakeholder distributes a total of 100 points ($, euro or coins) on the requirements, the Product Owner then will sum up the points and present the derived ordering of the requirements. Although the desired features will be selected at this point but there could be the chance that the selected feature will not provide benefit in terms of cost, time or easiness as much as it could have provided with other features selected at this time. The second step is Numerical assignment of requirements; it’s the most common technique for prioritizing requirements and is based on grouping requirements into different priority groups. For example group the requirements gathered from first steps into different groups based on their nature such as risk requirements, value requirements, and complex requirements etc. After this, requirements will be grouped based on influencing factors that could be effecting these requirements in any way. For example R1 and Rn are risk requirements [11] (see fig 2 below) and they are in any way contradicting with other requirements at the moment that have also been selected to implement in the sprint. Fig.1 depicts the steps of the proposed methodology.

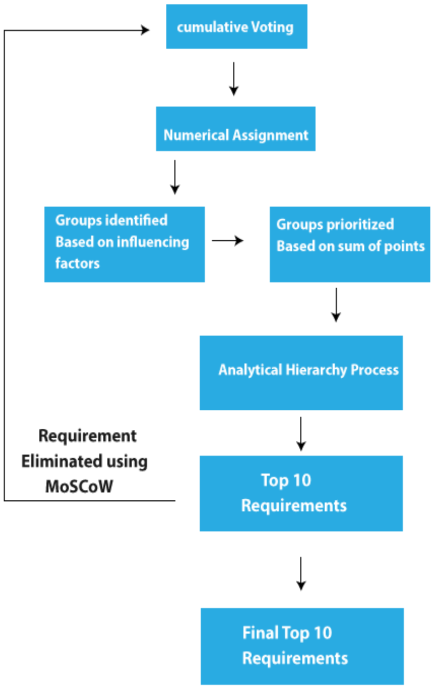

cumulative Voting
Numerical Assignment
Groups identified Based on influencing factors
Groups prioritized Based on sum of points
Analytical Hierarchy Process
Requirement Eliminated using MoSCoW
Top 10 Requirements
Final Top 10 Requirements

Fig. 1. Proposed hybrid model for requirements prioritization

This will cause trouble in implementing all of these requirements, therefore, it should be taken care of while selecting and prioritizing requirements for a sprint. Next, the groups will be prioritized based on highest points (see fig 2). Groups with requirements R1, R3, R4 have greater number of points as a whole then the other group therefore it has higher priority than other. After this the next step is AHP, in AHP the priorities of requirements is calculated to estimate their relative importance by comparing all unique pairs of requirements. In other words, the individual performing the comparison will decide manually which requirement has more significant, and to what extent using a scale 1-9.[14] AHP provides better results than any other tested methods as it is a ratio scale methodology, and also includes a consistency check.

Steps involved in AHP are:

1) Make an v×v matrix (v represents the number of requirements) requirements are latter inserted in rows and columns of the matrix.

2) For each pair of requirements, insert their relative intensity of importance (where the row of X meets the column Y). At the same point, insert the reciprocal values to the transposed positions (e.g. if cell XY=4 then cell YX=1/4)

3) Now, calculate the eigenvalues of this matrix to get the relative priority of each requirement. The final result will be the relative priorities of the requirements.

Total no. of comparisons that AHP requires is v×(v−1)/2. Redundancy is produced in pair-wise comparisons in AHP, therefore AHP also calculates the consistency ratio to check the accuracy of the comparisons [14].

At this point when small number of requirements have been selected and grouped, it is best to apply AHP at this point as grouping the requirements based on their nature and influencing factors will make it easy to check requirements with other groups and find out their relative importance, or contradiction between them. As Agile development team and PO have best idea because of their experience in the field about the implementation of such requirements that are conflicting each other to some extend and/or the risk or cost while implementing them it is suggested to apply MoSCoW at this point. MoSCoW is based on human opinion based on their experience, desire and influencing factors at that time such as market demand, cost, risk, time and resources, the resultant selected requirements are then again filtered using MoSCoW, this is expected to filter out those requirements that may have gotten higher points during the 100 dollar test (cumulative voting) but are causing contradiction to other requirements or may be less beneficial to get them implemented in this sprint. New requirements from the backlog are added after such requirements have been filtered out. If the number of newly added requirements is greater than 3 or 4 then all the steps are repeated on those newly added requirements. If small number of requirements is being added then only MoSCoW should be applied.

Detailed diagram of the Proposed Model is presented below.

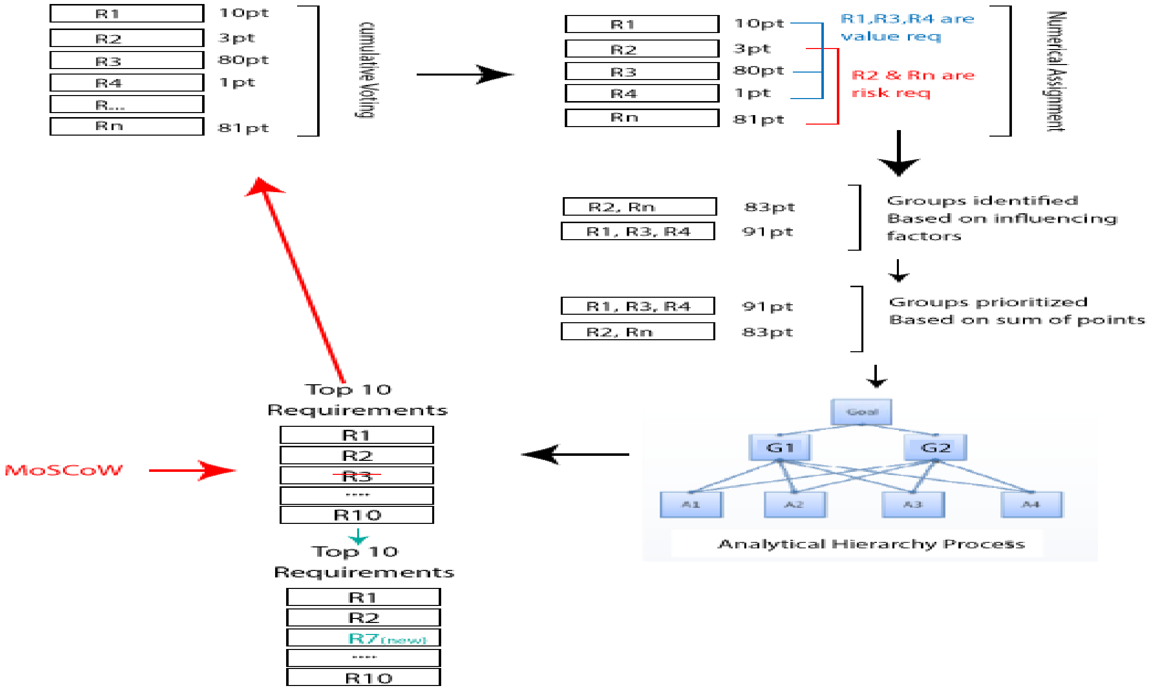


Fig. 2. Detailed presentation of the proposed model

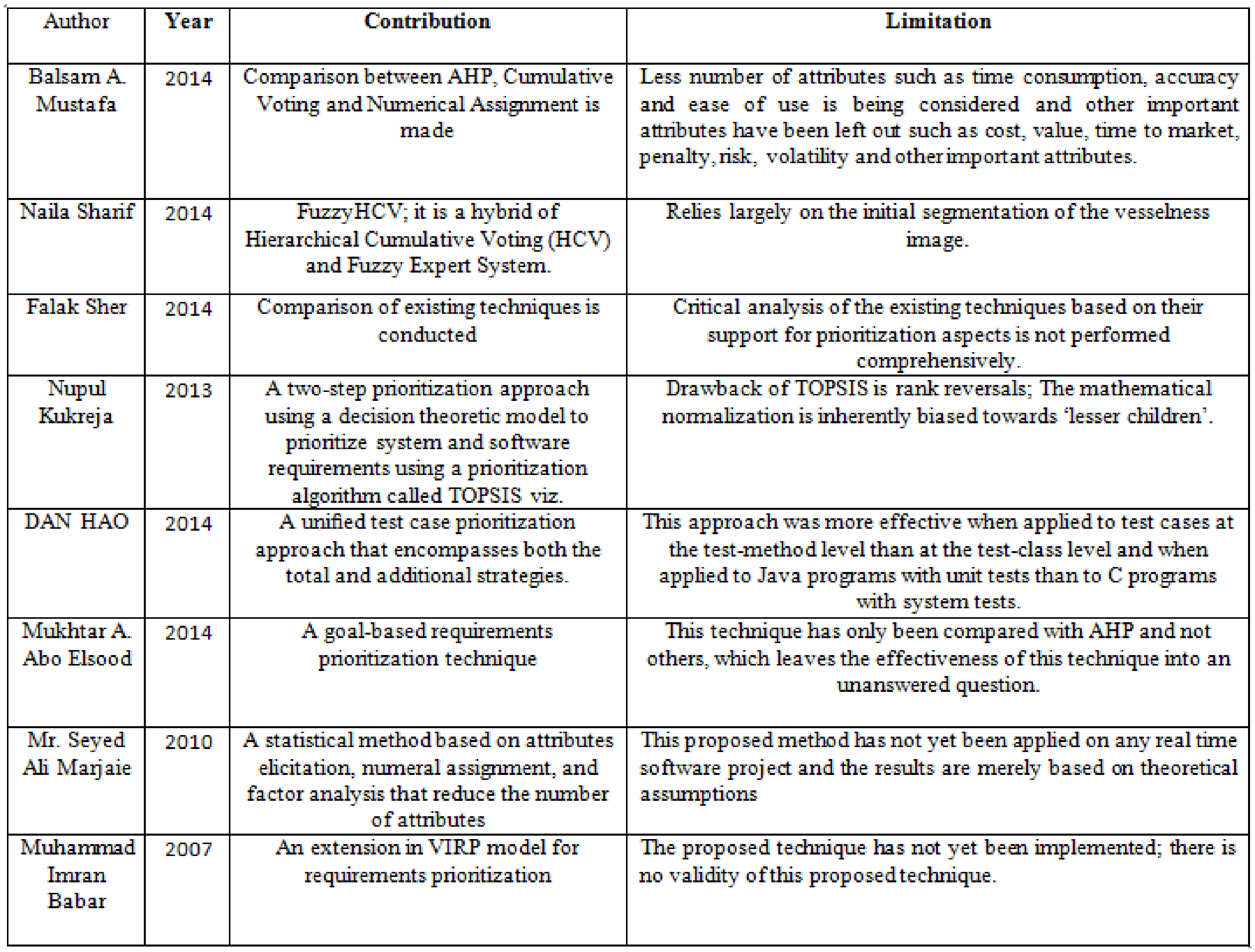
V. Findings & Assessments

The detail findings and assessments in this are comprehensively discussed and are evaluated in this section. We have devised a requirement prioritization framework to be considered in scrum model that can provide software requirements engineers/ testers encouraging feedback regarding adopting appropriate prioritization (validation) approach at a particular stage of software requirements process, the future direction could be the complete automation of software requirements process.

TABLE I. Comparison of Most Used Techniques

Technique	Scale	Granularity	Sophistication	Aspect	Perspective
AHP	Ratio	Fine	Very Complex	Strategic Importance, Penalty	Product Manager
100-Dollars Test	Ratio	Fine	Complex	Customer importance	Customers
Ranking	Ordinal	Medium	Easy	Volatility	Requirements Specialist
Numerical Assignment	Ordinal	Coarse	Very Easy	Time, Risk	Project Manager, Requirements Specialist
Top 10	---	Extremely Coarse	Extremely Easy	Customer importance	Customers

TABLE II. COMPARISON OF EXISTING WORK RELATED TO REQUIREMENTS PRIORITIZATION



Author	Year	Contribution	Limitation
Balsam A. Mustafa	2014	Comparison between AHP, Cumulative Voting and Numerical Assignment is made	Less number of attributes such as time consumption, accuracy and ease of use is being considered and other important attributes have been left out such as cost, value, time to market, penalty, risk, volatility and other important attributes.
Naila Sharif	2014	FuzzyHCV; it is a hybrid of Hierarchical Cumulative Voting (HCV) and Fuzzy Expert System.	Relies largely on the initial segmentation of the vesselness image.
Falak Sher	2014	Comparison of existing techniques is conducted	Critical analysis of the existing techniques based on their support for prioritization aspects is not performed comprehensively.
Nupul Kukreja	2013	A two-step prioritization approach using a decision theoretic model to prioritize system and software requirements using a prioritization algorithm called TOPSIS viz.	Drawback of TOPSIS is rank reversals; The mathematical normalization is inherently biased towards 'lesser children'.
DAN HAO	2014	A unified test case prioritization approach that encompasses both the total and additional strategies.	This approach was more effective when applied to test cases at the test-method level than at the test-class level and when applied to Java programs with unit tests than to C programs with system tests.
Mukhtar A. Abo Elsood	2014	A goal-based requirements prioritization technique	This technique has only been compared with AHP and not others, which leaves the effectiveness of this technique into an unanswered question.
Mr. Seyed Ali Marjaie	2010	A statistical method based on attributes elicitation, numeral assignment, and factor analysis that reduce the number of attributes	This proposed method has not yet been applied on any real time software project and the results are merely based on theoretical assumptions
Muhammad Imran Babar	2007	An extension in VIRP model for requirements prioritization	The proposed technique has not yet been implemented; there is no validity of this proposed technique.

TABLE III. COMPARISON TABLE OF TECHNIQUES IN TERMS OF TECHNICAL AND BUSINESS ASPECTS

No.	Techniques	Technical Aspects						Business/Client Aspects			
		Citations	Scalability	Ease of Use	Time Complexity	Decision making	Accuracy	Sales	Marketing	Customer Satisfaction	Strategic
1	Analytic Hierarchy Process (AHP)	50		Yes		Yes	Yes				
2	Binary-Tree Prioritize	8		Yes			Yes				
3	Bubble Sort	8		Yes		Yes				Yes	
4	Cumulative voting (CV)	20		Yes	Yes		Yes				
5	Kano Analysis	5		Yes		Yes	Yes		Yes	Yes	
6	MoSCoW	6	Yes	Yes							Yes
7	Pair-wise analysis	10		Yes	Yes	Yes	Yes				
8	Numeral Assignment	15		Yes	Yes	Yes	Yes				
9	Ranking	8	Yes	Yes	Yes	Yes	Yes	Yes			
10	Relative weighting	2		Yes							
11	Top Ten Requirements	18		Yes	Yes	Yes	Yes				
12	Wieger's Prioritization	14	Yes	Yes	Yes	Yes	Yes			Yes	

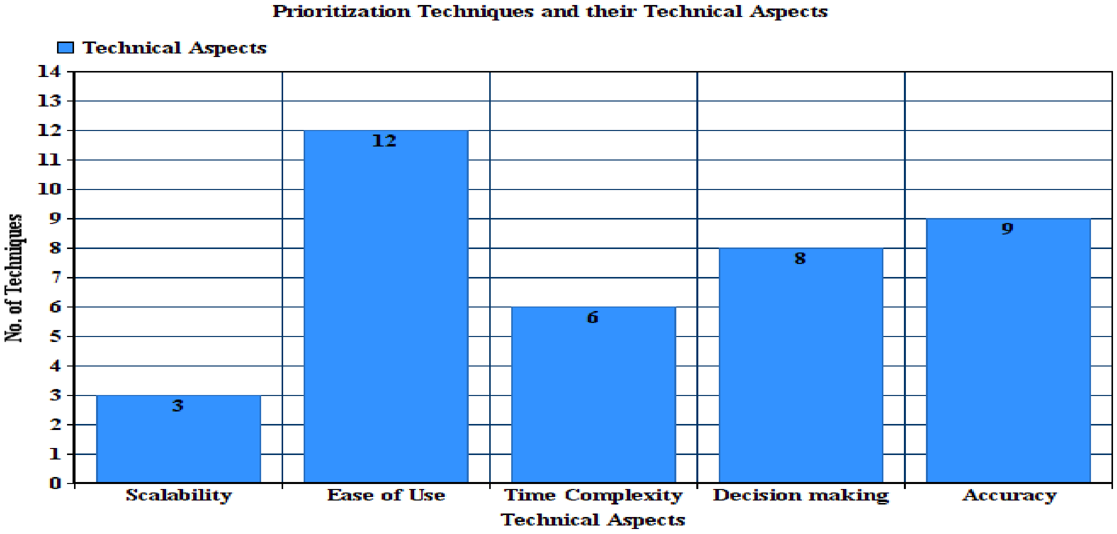


Fig. 3. A detail representation between requirements prioritization models and corresponding quality attributes (Technical Aspect)

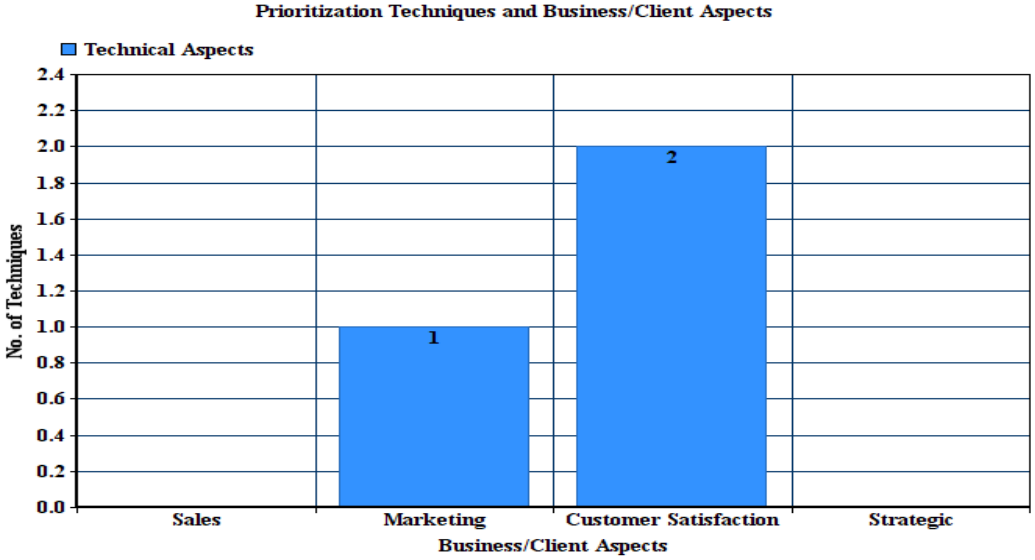


Fig. 4. A detail representation between requirements prioritization techniques and Business Client Aspect

VI. Systematic Methodology

Systematic literature review (SLR) has been done in order to discover new findings. Based on different research problems (mentioned in the literature) that are associated to both different prioritization aspects and techniques; we found the motivation for this research and hence the research questions are designed accordingly.

VII. Conclusion

As requirements emerge throughout the software development process and are needed to be prioritized and managed with highest priority, especially in the case of Agile Software Development process. As disused and highlighted in this research work, there are many requirements prioritization techniques, methodologies proposed and been followed but most of them fail to take account of all those factors that play an important role in prioritizing requirements and in overall quality of software product being developed. After a comprehensive literature, it is found that the existing prioritization techniques do not span over all type of projects. Some techniques work well on agile development process and some on traditional development. Therefore there is a need of a prioritization technique that considers the above mentioned factors (time factor and human behavioral factor) while prioritizing requirements.

References

[1] Elsevier (2012) A decade of agile methodologies: Towards explaining agile software development, The Journal of Systems and Software

[2] Mohd. Muqeem, Dr.Mohd.Rizwan, Validation of Requirement Elicitation Framework using Finite State Machine", IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp 1210 – 1216, 2014.

[3] Ming Huo, June Verner, Liming Zhu, Muhammad Ali Babar (2004) Software Quality and Agile Methods, IEEE

[4] Lan Cao, Balasubramaniam Ramesh (2008) Agile Requirements Engineering Practices:An Empirical Study, IEEE.

[5] Shahid Nazir, SEN-2005, Why Quality? ISO 9126 Software Quality Metrics (Functionality) Support by UML Suite, NY, USA.

[6] DOI= 1050849.1050860

[7] Pekka Abrahamssona, Juhani Warstab, Mikko T. Siponenb and Jussi Ronkainen (2003) New Directions on Agile Methods: A Comparative Analysis, IEEE.

[8] Armin Eberlein, Julio Cesar Sampaio do Prado Leite (2002) Agile Requirements Definition: A View from Requirements Engineering, Proceedings of the International Workshop on Requirement engineering.

[9] S. N. Bhatti, Deducing the complexity to quality of a system using UML. ACM SIGSOFT Software Engineering Notes 34(3): 1-7 (2009). DOI=1527202.1527207

[10] DAN HAO, LINGMING ZHANG, LU ZHANG, GREGG ROTHERMEL, HONG MEI, (2014) A Unified Test Case Prioritization Approach, ACM Transactions on Software Engineering and Methodology, Vol. 24, No. 2, Article 10, Pub. date: December 2014.

[11] Frauke Paetsch, Frauke Paetsch, Dr. Frank Maurer (2003) Requirements Engineering and Agile Software Development, IEEE

[12] V. N. Vithana (2015) Scrum Requirements Engineering Practices and Challenges in Offshore Software Development, International Journal of Computer Applications (0975 – 8887), Volume 116 – No. 22, April 2015.

[13] Azar, J.,Smith, R.K., "Value-Oriented Requirements Prioritization in a Small Development Organization", IEEE Computer society, 2007, pp 32 – 37, 2007.

[14] Anna Perini , Angelo Susi , Paolo Avesani (2013) A Machine Learning Approach to Software Requirements Prioritization, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. 4, APRIL 2013

[15] Rahul Thakurta (2013) A framework for prioritization of quality requirements for inclusion in a software project, Software Quality Journal (2013) 21:573–597

[16] Naila Sharif, Kashif Zafar, Waqas Zyad (2014) Optimization of Requirement Prioritization using Computational Intelligence Technique, 2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (iCREATE) Islamabad, Pakistan, April 22-24, 2014

[17] Nupul Kukreja, Barry Boehm (2013) Integrating Collaborative Requirements Negotiation and Prioritization Processes: A Match Made in Heaven, Proceedings of the 2013 International Conference on Software and System Process

[18] Rubaida Easmin, Alim Ul Gias, Shah Mostafa Khaled (2014) A Partial Order Assimilation Approach for Software Requirements Prioritization 3rd INTERNATIONAL CONFERENCE ON INFORMATICS, ELECTRONICS & VISION 2014

[19] Shahid N. Bhatti, Maria Usman, Amr A. Jadi, 2015, Validation to the Requirement Elicitation Framework via Metrics. ACM SIGSOFT Software Engineering Notes 40(5): 17, USA. DOI= 2815021.2815031

[20] J. Karlsson and K. Ryan. 1997, "Prioritizing requirements using a cost-value approach," IEEE Software 14 (5), pp. 67–74.

[21] John A Mcdermid, Software Engineer's Reference Book, Butterworth-Heinemann, 1991.

[22] Muhammad Ramzan, M. Arfan Jaffar and Arshad Ali Shahid (2011) VALUE BASED INTELLIGENT REQUIREMENT PRIORITIZATION (VIRP): EXPERT DRIVEN FUZZY LOGIC BASED PRIORITIZATION TECHNIQUE, International Journal of Innovative Computing, Information and Control, Volume 7, Number 3, March 2011.

[23] Mohd. Sadiq, Jawed Ahmed, Mohammad Asim, Aslam Qureshi , R. Suman (2010) More on Elicitation of Software Requirements and Prioritization using AHP, 2010 International Conference on Data Storage and Data Engineering

[24] M. Waseem Asghar, Alessandro Marchetto, and Angelo Susi Fondazione Bruno Kessler , Giuseppe Scanniello (2013) Maintainability-based Requirements Prioritization by using Artifacts Traceability and Code Metrics, 2013 17th European Conference on Software Maintenance and Reengineering

[25] Richard Berntsson Svensson, Tony Gorschek, Björn Regnell, Richard Torkar, Ali Shahrokn, Robert Feldt, Aybuke Aurum (2011) Prioritization of Quality Requirements: State of Practice in Eleven Companies, 2011 IEEE 19th International Requirements Engineering Conference

[26] Mukhtar A. Abo Elsood, Hesham A. Hefny , Eman S. Nasr (2014) A Goal-Based Technique for Requirements Prioritization, The 9th International Conference on INFOrmatics and Systems (INFOS2014) - 15-17 December Software Engineering - Challenges of Openness Track

[27] Nikita Garg , Dr. Pankaj Agarwal , Shadab Khan (2015) Recent Advancements in Requirement Elicitation and Prioritization Techniques, 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India

[28] Mr. Seyed Ali Marjaie , Mrs. Vasundhara Kulkarni, 'Recognition of Hidden Factors In Requirements Prioritization Using Factor Analysis', IEEE

[29] Muhammad Imran Babar, Muhammad Rarnzan, Shahbaz A. K. Ghayyur (2007) Challenges and Future Trends in Software Requirements Prioritization, IEEE

[30] Muhammad Aasem, Muhammad Ramzan and Arfan Jaffar, 'Analysis and optimization of software requirements prioritization techniques', IEEE

A Novel Structure of Advance Encryption Standard with 3-Dimensional Dynamic S-box and Key Generation Matrix

Ziaur Rahaman[1], Anjela Diana corraya[2], Mousumi Akter Sumi[3], Ali Newaz Bahar[4]
Department of Information and Communication Technology
Mawlana Bhashani Science and Technology University
Tangail, Bangladesh

***Abstract*—The study of sending and receiving secret messages is called cryptography. Generally, senders and receivers are unaware about the process of encryption and decryption. Hence, encryption plays an important role in data communication and data security. The meaning of encryption is not only to keep data confidential from unwanted access but also ensuring the data integrity through available way. As the capacity of breaking the security is increasing rapidly, so, the process that hides information is one of the most concerned topics. Advanced Encryption Standard is a popular, widely used and efficient encryption algorithm, which has been used since it was invented. This paper focuses on the AES key generation process and Substitution box. It modifies the conventional key generation technique and builds the dynamic 3-Dimensional S-box of Advance Encryption Standard. The proposed approach suggests 3-Dimensioanl Key Generation Matrix and S-box. As per shown this novel technique increases the amount of time it needs during encryption and decryption. The experimental result shows that it also enhances the strength of AES algorithm. The proposed approach illustrates the theoretical analysis and corresponding experimented results.**

Keywords*—Advanced Encryption Standard; AES Modification; 3-dimensional Key Generation Matrix; dynamic S-box

I. Introduction

The data transmission rate over the Internet has been getting massively increased. So in order to give full assurance over secured data transmission from sender to receiver is a great concern in this universe. Besides the confidentiality, data integrity is another important issue. Advanced Encryption Standard (AES) plays a vital rule to insure the data integrity and confidentiality. Rijndael is the original name of AES [1][2] which is established by National Institute of Standards and Technologies (NIST)[3]. Ciphers family includes different key and block sizes belongs to Rijndael [4]. AES [5] is a block cipher system divided into Diffusion and Confusion principles. In confusion, the length of plaintext and cipher text is same. But in diffusion the length of plaintext and cipher text is unequal. In enciphering system, key is the unavoidable part. So, in this paper, we have proposed a new modified key scheduling algorithm. On the other hand, AES is based on the S-box that increases the cryptographic strength. For this, we have generated dynamic 3-Dimensional S-box. In cryptanalysis, we know that Advanced Encryption Standard (AES) is generated from Galois Field GF (28) and introduced AES-128, AES-192 and AES-256. With observation it is clear that AES-192 is slower than AES-128 but AES-256 is more secure than AES 128. AES-256 is used to protect against quantum brute force attack.

The objective of this paper is to work with GF (35) which is feasible for 243 bits plaintext and 243 bits keys at a time. In order to modify traditional AES and to make it more efficient, proposed system has involved a 3-Dimensional Key Generation Matrix (3DKGM) for key generation and 3-Dimensional Dynamic S-box. The total number of round in this system is 16 which is divided into two parts named odd round (1,3,5,,15) and even round (2,4,6,,16). The main difference of these two rounds is absence of mix column in odd round and present in even round. So, it enhances complexity as well as complexity expensive for hackers. As a result, the proposed system stands as a secured system.

The rest of this paper is organized as follows: Issues and security in section 2, Related works in section 3, Describe problem statement in section 4, section 5 and 6 respectively go for 3-Dimensional Key Generation Matrix (3DKGM) System and Proposed 3-Dimensional Dynamic S-box, Section 7 represents the Proposed system, All experimental analysis and discussion take place in section 8 and finally Future perspective and conclusion in section 9.

II. Issues in Security

The three security issues are: confidentiality, integrity and availability; known as the ACI triad [6].

A. Availability

The information that is formed and stored needs to be available to the authorized users. Without availability, the information is useless.

B. Integrity

Integrity assures that the information is changed by authorized entities and through authorized mechanisms. Unwanted change in information occurs risk to integrity. The sent data must be same as received data and not be altered through transmission path.

C. Confidentiality

Data theft and unauthenticated access are raised [7] which can be protected by confidentiality. Confidentiality means to guard against danger that is ensured by data encryption services, authentication and security protocols.

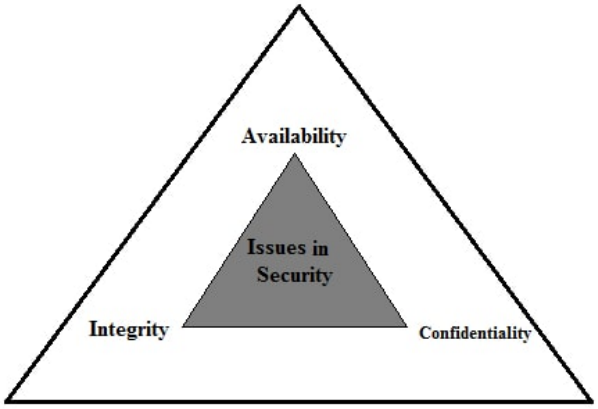


Fig. 1. Issues in Security

III. Related Works

In this paper, we will just concentrate on the core components of AES key scheduling and AES S-box. Modification of these two properties is carried out in this paper. All the parts of modifications over AES is described here to make understand the differences of our proposed AES with traditional AES.

In 1997, a great deal of corporation is attack and in 1998 for implementation machines are too expensive. After that, with time computers power is increasing and as a result there required stronger algorithm to face hackers attacks. Various works which are done earlier in AES field by many researchers. We analysis their work, idea, platform, limitations and also draw a survey in this section. In [5], they represent ASIC AES implementations of power analysis against the attack and extract information on side channel attack. But, at the time of simulating attack a huge number of noises are present. In [8], showed an image encryption algorithm for high definition image this is based on the modification of AES. But, it has an unstable round number at the time of attack and also the long time recruitment s for the encryption and decryption process. In [9], they proposed an improved algebraic expression in the S-box generation which made the generation process more complicated. But, the limitation, excesses the computational cost over the improvement. In [6], several security issues are described that are concerned for cloud computing and also showed a number of serious security threats. In [10], a new system is proposed for data security called RSA. RSA algorithm is used here to encrypting a large database or store data into any files. But limitation was this system is better for static data, but not better for linear methods with the retrieval speed.

For the experimental analysis of our key generation, a number of different file sizes are used to show their computational time and for the S-Box, total time is calculated for different number orientation. So with the analysis of result we can show that our approach is more efficient compared with other algorithm. The proposed system in this paper can provide a perfect combination of excellent security, efficiency, flexibility, implement-ability and performance.

IV. Problem Statement

AES based on Rijndel Algorithm which is a standard combination of a strong algorithm and a strong key. With the calculation it can see that for AES-128 to check all possible key (50 billion keys per seconds) total required time is 5*1021 years [11]. as the steps of the proposed algorithm in this paper is more complex than traditional AES, so, the total required time is obviously more than the traditional one.

There are two options to ensure the security rate of any algorithm. These are based on time and cost. If an intruders required time to break a system is greater than his life time, then that system is called a secure system.

And again if the total cost that required breaking the system is much more then its initial making cost. Then the system is called secure over cost. Because for any intruder its obviously not a choice to break the system which belongs much more cost then its initial rate. With the observation of above drawbacks of traditional AES algorithm, this paper introduced a system to replace the 2-Dimensional process to 3-Dimensional process. This 3-Dimensional is used not only for key generation but also for S-box to make it dynamic. The step by step process of this system is capable for disarranging of initial message and key, that is enough to confuse the intruder. And this is happened by occurring bit operation complexity which leads to hackers drifting into undecidable problems.

V. 3-Dimensional Key Generation Matrix (3DKGM)

For the modified AES key generation, we used a 999 cube matrix which is shown at the Fig. 2. With this system, we can overcome the limitation to calculate 35. Our 3D matrix is a combination of Latin Alphabets (A-Z), integer value (0-9) and Greek symbols. So it can make the secrete massage more and more complex. The overview of this matrix can be seen on Fig. 2. To better understand, let consider a secret key: COMPLEX@+ , which have to encrypt. At first of the process, the position of every byte is to be declared. In Table 1 the position for secrete key COMPLEX@+is declare d.

A. Encryption process in 3-Dimensional Key Generation Matrix System

To understand the encryption process of proposed 3DKGM system, we have shown an example in below. Here, it can be seen that from Fig. 3 for P, the row is first found (at x-axis) and the column (at y-axis) number from the 3D matrix.

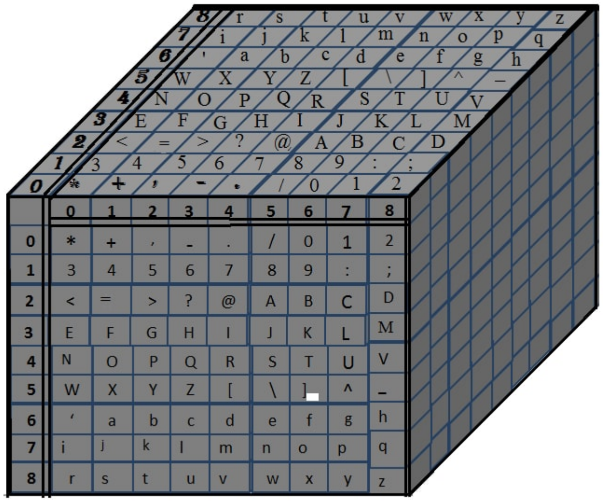
TABLE I. Position of Secret Key

Secrete Message	C	O	M	P	L	X	@	+	α	μ
Position	0	1	2	3	4	5	6	7	8	0

After that, at the z-axis we find the position of letter P. So for P we can write 42G. So we get,

$$P = 42G \tag{1}$$

Here, 4 for row, 2 for column and G for position at 3 (can see from the Table 1).



	0	1	2	3	4	5	6	7	8
0	*	+	,	-	.	/	0	1	2
1	3	4	5	6	7	8	9	:	;
2	<	=	>	?	@	A	B	C	D
3	E	F	G	H	I	J	K	L	M
4	N	O	P	Q	R	S	T	U	V
5	W	X	Y	Z	[	\	]	^	_
6	‘	a	b	c	d	e	f	g	h
7	i	j	k	l	m	n	o	p	q
8	r	s	t	u	v	w	x	y	z

Fig. 2. Extended 3D Polybius Cube Matrix

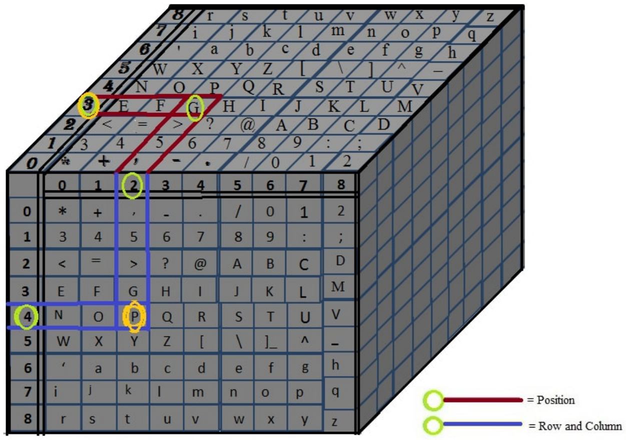


Fig. 3. Encryption procedure of ”P”

With this encryption process the original text of key will be encrypted with some logical codes which is very much harder to identify. Here, the plain text is P and the encrypted output for this 16G.

VI. Proposed 3-Dimensional Dynamic S-Box

As discussed earlier, a dynamic 3-Dimensional S-box is generated. First, an initial S-box is needed which is 3-Dimensional and based on hexadecimal numbers. Fig. 4 and Fig. 6 are our generated initial 3-Dimensional S-box.

The 3-dimensional S-box is defined like:
X (a, b, c) = Y1, Y2, Y3 Where,
a = index of x-axis

b = index of y-axis

c = index of z-axis

Y1 = Value of the row

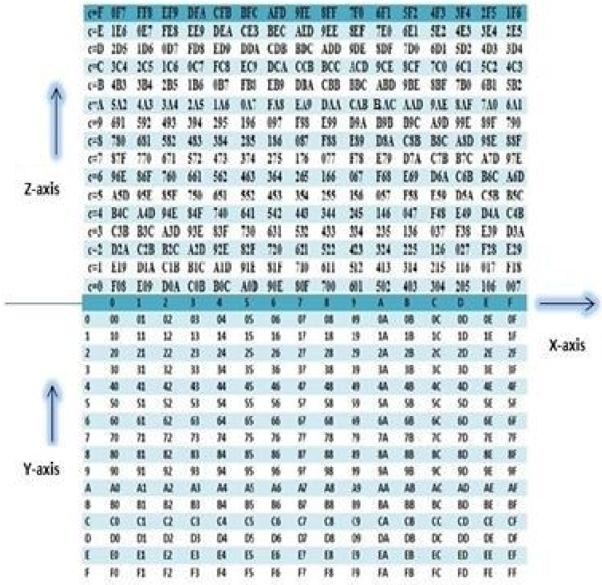
Y2 = Value of the column

Y3 = Index value of c after selecting (Y1, Y2)

= (1st element of reversing hexadecimal number (next c follows the one byte cyclic left rotation of numbers), c, D) after selecting (Y1, Y2)

D = Hexadecimal number starts from (16/2+1)th value and run up to last and again in a cyclic order starts from the first to (16/2)th value (follows the number cycling).

For better understanding, let us take an example: the XOR-ed result of plaintext and key is ABC. Now, from the above Fig. 4, For A we go through the x-axis and then y-axis for B. Finally, for C we go through z-axis and select the value corresponding x, y and z-axis which is BCC. In traditional S-box, hardware implementation is too hard. But, in proposed S-box, we can avoid these complexities.


Z-axis
Y-axis
X-axis

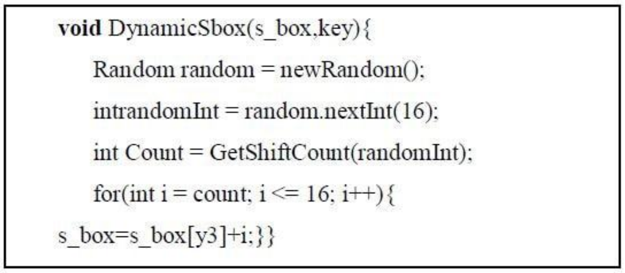
c=F	0F7	FF8	EF9	DFA	CFB	BFC	AFD	9FE	8FF	7F0	6F1	5F2	4F3	3F4	2F5	1F6
c=E	1E6	0E7	FE8	EE9	DEA	CEB	BEC	AED	9EE	8EF	7E0	6E1	5E2	4E3	3E4	2E5
c=D	2D5	1D6	0D7	FD8	ED9	DDA	CDB	BDC	ADD	9DE	8DF	7D0	6D1	5D2	4D3	3D4
c=C	3C4	2C5	1C6	0C7	FC8	EC9	DCA	CCB	BCC	ACD	9CE	8CF	7C0	6C1	5C2	4C3
c=B	4B3	3B4	2B5	1B6	0B7	FB8	EB9	DBA	CBB	BBC	ABD	9BE	8BF	7B0	6B1	5B2
c=A	5A2	4A3	3A4	2A5	1A6	0A7	FA8	EA9	DAA	CAB	BAC	AAD	9AE	8AF	7A0	6A1
c=9	691	592	493	394	295	196	097	F98	E99	D9A	C9B	B9C	A9D	99E	89F	790
c=8	780	681	582	483	384	285	186	087	F88	E89	D8A	C8B	B8C	A8D	98E	88F
c=7	87F	770	671	572	473	374	275	176	077	F78	E79	D7A	C7B	B7C	A7D	97E
c=6	96E	86F	760	661	562	463	364	265	166	067	F68	E69	D6A	C6B	B6C	A6D
c=5	A5D	95E	85F	750	651	552	453	354	255	156	057	F58	E59	D5A	C5B	B5C
c=4	B4C	A4D	94E	84F	740	641	542	443	344	245	146	047	F48	E49	D4A	C4B
c=3	C3B	B3C	A3D	93E	83F	730	631	532	433	334	235	136	037	F38	E39	D3A
c=2	D2A	C2B	B2C	A2D	92E	82F	720	621	522	423	324	225	126	027	F28	E29
c=1	E19	D1A	C1B	B1C	A1D	91E	81F	710	611	512	413	314	215	116	017	F18
c=0	F08	E09	D0A	C0B	B0C	A0D	90E	80F	700	601	502	403	304	205	106	007
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
1	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
2	20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
3	30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
4	40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
5	50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
6	60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
7	70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
8	80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
9	90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

Fig. 4. 3-Dimensional S-box

A. DynamicSbox Algorithm

A word called dynamic is also used for 3-dimensional S-box. So, after getting the total initial keys, a random number will be generated. According to the random number, the keys will be left-rotated. The value of rotation will be stored in a variable and according to this value; the S-box will be rotated again. Fig. 5 is showing an algorithm of dynamic S-box. For example, if the Count =2 (rotation is 2 times for keys), then the value of dynamic 3-Dimensional S-box is given in Fig. 7.

The random numbers and the rotation of S-box depend on the number 16 as the total number of hexadecimal is 16. If the shifting value in each z-axis is greater than 16, then no rotation is occurred in initial 3-Dimensional S-box. The rotation will occur in z-axis only as it is the main part of dynamic 3-Dimensional S-box.



```
void DynamicSbox(s_box,key){
    Random random = newRandom();
    intrandomInt = random.nextInt(16);
    int Count = GetShiftCount(randomInt);
    for(int i = count; i <= 16; i++){
s_box=s_box[y3]+i;}}
```

Fig. 5. An algotithm of dynamic S-box

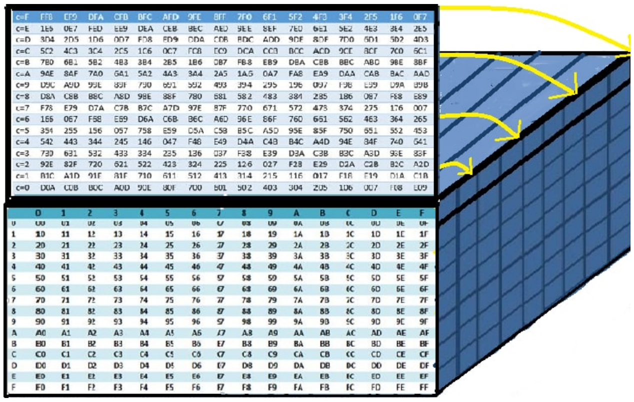
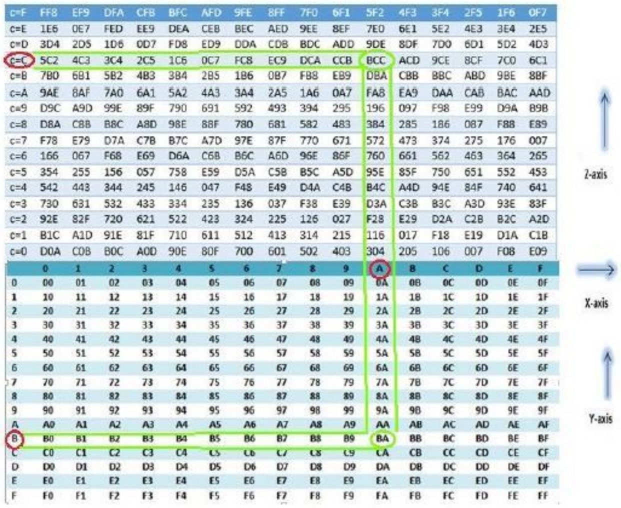


Fig. 6. 3-Dimensional S-box

VII. Proposed System

The Science of information transmitting and retrieving securely over the insecure channel is called Cryptography [12]. There are two parts in the study of cryptography. They are: encryption and decryption. The 1st part is Encryption where a sender is converting data into an unintelligible string or cipher text during transmission, so that a hacker could not know about the sent data. And the 2nd part ,Decryption is just the reverse of it. With a proper decryption process the receiver converts senders cipher text into a plaintext [13], as a meaningful text. In the AES algorithm, the 1st step is the XOR operation of plaintext and the key of same length of bits [3]. So in this work, we have focused on the key generation method and next is an S - box. We know, the background of AES depends on Galois Field. The working procedure of the system is given as follow:

- To follow the rules of Galois Field, we have to take (P)n number of bits as a key and as a plaintext. Where, P has to be a prime number and n to be any integer value. So we used here 243 bits (35). But to convert these 243 bits into bytes (hexadecimal), we perform zero padding (5 zeros) at the last place of the key. Total 248 bits are obtained here (243+5=248); now we convert the bits into byte and gets 31 bytes (2488=31).
- In the next step, the proposed 3DKGM is used and 3 byte is obtained for every 1 byte. Details are described about our proposed 3-Dymensional Key Generation matrix system in the previous section. So, after using the logical calculations of 3-Dimensional matrix, we get 93 bytes (313=93) means 744 bits.



c=F	FF8	EF9	DFA	CFB	BFC	AFD	9FE	8FF	7F0	6F1	5F2	4F3	3F4	2F5	1F6	0F7
c=E	1E6	0E7	FED	EE9	DEA	CEB	BEC	AED	9EE	8EF	7E0	6E1	5E2	4E3	3E4	2E5
c=D	3D4	2D5	1D6	0D7	FD8	ED9	DDA	CDB	BDC	ADD	9DE	8DF	7D0	6D1	5D2	4D3
c=C	5C2	4C3	3C4	2C5	1C6	0C7	FC8	EC9	DCA	CCB	BCC	ACD	9CE	8CF	7C0	6C1
c=B	7B0	6B1	5B2	4B3	3B4	2B5	1B6	0B7	FB8	EB9	DBA	CBB	BBC	ABD	9BE	8BF
c=A	9AE	8AF	7A0	6A1	5A2	4A3	3A4	2A5	1A6	0A7	FA8	EA9	DAA	CAB	BAC	AAD
c=9	D9C	A9D	99E	89F	790	691	592	493	394	295	196	097	F98	E99	D9A	B9B
c=8	D8A	C8B	B8C	A8D	98E	88F	780	681	582	483	384	285	186	087	F88	E89
c=7	F78	E79	D7A	C7B	B7C	A7D	97E	87F	770	671	572	473	374	275	176	007
c=6	166	067	F68	E69	D6A	C6B	B6C	A6D	96E	86F	760	661	562	463	364	265
c=5	354	255	156	057	758	E59	D5A	C5B	B5C	A5D	95E	85F	750	651	552	453
c=4	542	443	344	245	146	047	F48	E49	D4A	C4B	B4C	A4D	94E	84F	740	641
c=3	730	631	532	433	334	235	136	037	F38	E39	D3A	C3B	B3C	A3D	93E	83F
c=2	92E	82F	720	621	522	423	324	225	126	027	F28	E29	D2A	C2B	B2C	A2D
c=1	B1C	A1D	91E	81F	710	611	512	413	314	215	116	017	F18	E19	D1A	C1B
c=0	D0A	C0B	B0C	A0D	90E	80F	700	601	502	403	304	205	106	007	F08	E09

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
1	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
2	20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
3	30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
4	40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
5	50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
6	60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
7	70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
8	80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
9	90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

Fig. 7. 3-DimensionalS-box when rotation value is 2

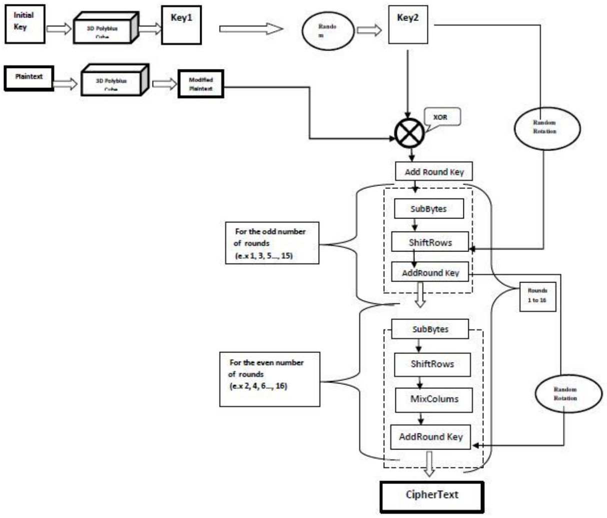


Fig. 8. The Modified AES Encryption Algorithm

- Now, to make these 744 bits more complex and also for matching with the proposed S-box, used random rotation method. In random rotation the bit will be rotated towards left at anti-clockwise.
- For plaintext, the same procedures used for transforming 243 (35) bits into 744 bits. After this step, we have XOR-ed these (744 bits) key with the Plaintext (744 bits). And the XOR resulted will be forwarded for next steps where S-box exists.
- The resulted bits will be forwarded to the round procedures. That is already mentioned above sections (16 rounds). These 16 rounds are partitioned into odd

round and even round.

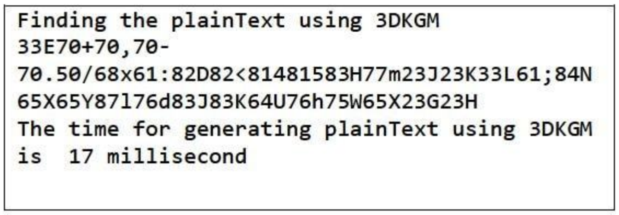
- Even round procedure includes the SubBytes , ShiftRows, MixCloums, AddRound Key.
- Odd round procedure includes the SubBytes, ShiftRows, AddRound Key.

The main different between the two round procedure is absent of Mixcloumns in the odd round.

It is already known that key generation process never create any bad impacts over the hardware implementation. In order to increase strength of key, the key generation process was tried to make harder and also S-box exists in SubBytes. In the Fig. 8, we can see the overall schema of our proposed modified AES algorithm.

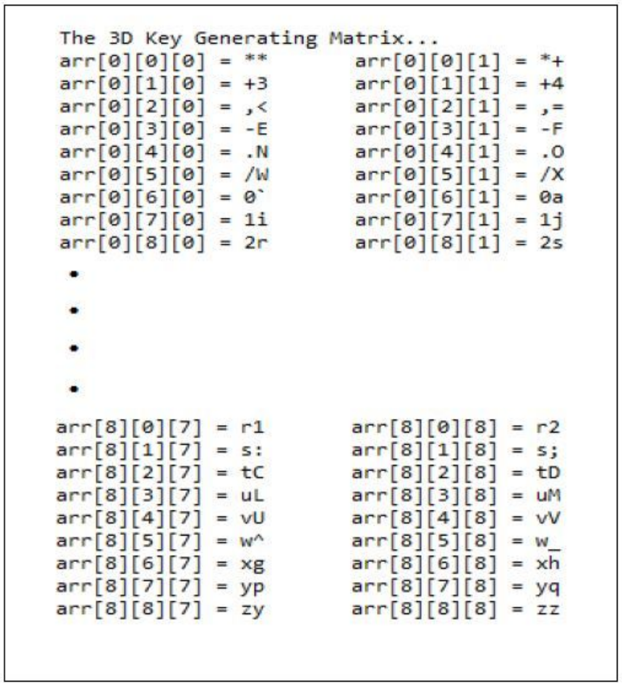
A. The Output -Code Snippet for 3-Dynamices Key Generation Matrix (3DKGM)Algorithm

Here in Fig. 9, Fig. 10 and Fig. 11, we can see some parts of the output for our key generation process using Java. In Fig. 9 we see the output for the message Information and Communication Technology using 3DKGM algorithm, which is our plaintext in the system. In Fig. 10, we can see some output matrix parts of the 3DKGM algorithm. As the length of the total matrix is so long, so we skip the middle parts in this paper for proper representation.



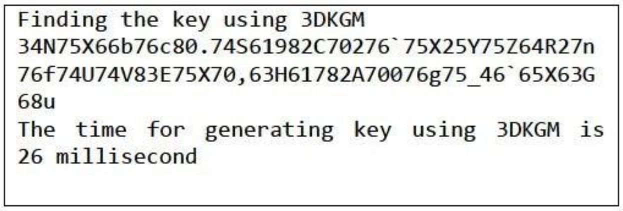
```
Finding the plainText using 3DKGM
33E70+70,70-
70.50/68x61:82D82<81481583H77m23J23K33L61;84N
65X65Y87l76d83J83K64U76h75W65X23G23H
The time for generating plainText using 3DKGM
is  17 millisecond
```

Fig. 9. Output for the PlainText using 3DKGM



```
The 3D Key Generating Matrix...
arr[0][0][0] = **        arr[0][0][1] = *+
arr[0][1][0] = +3        arr[0][1][1] = +4
arr[0][2][0] = ,<        arr[0][2][1] = ,=
arr[0][3][0] = -E        arr[0][3][1] = -F
arr[0][4][0] = .N        arr[0][4][1] = .O
arr[0][5][0] = /W        arr[0][5][1] = /X
arr[0][6][0] = 0`        arr[0][6][1] = 0a
arr[0][7][0] = 1i        arr[0][7][1] = 1j
arr[0][8][0] = 2r        arr[0][8][1] = 2s
 •

 •

 •

 •

arr[8][0][7] = r1        arr[8][0][8] = r2
arr[8][1][7] = s:        arr[8][1][8] = s;
arr[8][2][7] = tC        arr[8][2][8] = tD
arr[8][3][7] = uL        arr[8][3][8] = uM
arr[8][4][7] = vU        arr[8][4][8] = vV
arr[8][5][7] = w^        arr[8][5][8] = w_
arr[8][6][7] = xg        arr[8][6][8] = xh
arr[8][7][7] = yp        arr[8][7][8] = yq
arr[8][8][7] = zy        arr[8][8][8] = zz
```

Fig. 10. Output for the 3D key generating Matrix



```
Finding the key using 3DKGM
34N75X66b76c80.74S61982C70276`75X25Y75Z64R27n
76f74U74V83E75X70,63H61782A70076g75_46`65X63G
68u
The time for generating key using 3DKGM is
26 millisecond
```

Fig. 11. Output for key using 3DKGM

In Fig.11, here is shown the output for finding key using the 3DKGM algorithm.

After all these process, our next task is the rotation method. These step by step processes are shown before in the flow diagram on Fig. 8.

Form the Fig 8, we can see that next we perform the XOR-operation between the key and the modified Plaintext in our system. Then we are gone for further tasks, those belongs to the 3-Dymentional S-box.

VIII. Experimental Analysis and Discussion

Time to encrypt and decrypt is an important feature of any encryption algorithm. As, Key scheduling and S-box are the parts of encryption algorithm, so, time has great impact on these.

As because of dynamic 3-dimensional S-box, the strength of our proposed algorithm is increased. In security analysis, it will take too long time for brute force approach. Wadi and Zainal recently proposed a S-box based on modified AES-128 [3] block cipher which is too easy to break . We did two types of experiment and simulated on Matlab2010a and we solve our algorithm with Java.

A. Computational Time vs. File Size

In this part we showed a comparison result of our proposed approach with others 3 algorithms (AES, DES and TDES). We used different size of file and calculate the computational time. We know that the computational time for encryption process is the total time for the algorithm to convert plaintext into cipher text. In we can calculate the performance by calculation different time take by different algorithm. Here from the table II, we can see that, we take different file size of 20, 35, 155, 333 and 512.

And for 20kb we get the execution time 26, 25, 27, and 28 for the AES, DES, TDES and our proposed algorithm. Similarly, for 333Kb file size it gives 469,481, 509 and 501 for the AES, DES, TDES and our proposed algorithm.

So from the following table, we can see that, our proposed

TABLE II. COMPUTATIONAL TIME FOR ENCRYPTION IN SEC. VS FILE SIZE IN KB

File Size in KBl	AES	DES	TDES	Our Proposed
20	26	25	27	28
35	53	56	64	58
155	251	270	287	261
100	436	448	476	468
300	436	448	476	468
512	469	481	509	501

approach decrease the computational time rather than others algorithm. As a result it cans build up more security for the vast system.

From the table II, we draw a graph in order to show the vast dependency of computational time over file size. So from the above table, we showed the graphical representation in Fig. 12. Here it can be seen that the proposed approach has lower computational time compared to other algorithms for sophisticated bits. So finally we can say that our proposed algorithm is efficient than others algorithm from the comparison results of Table II and graph.

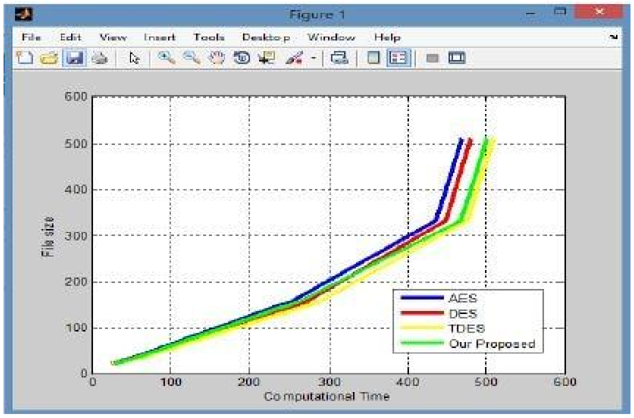


Fig. 12. Computational Time for Encryption in Sec.

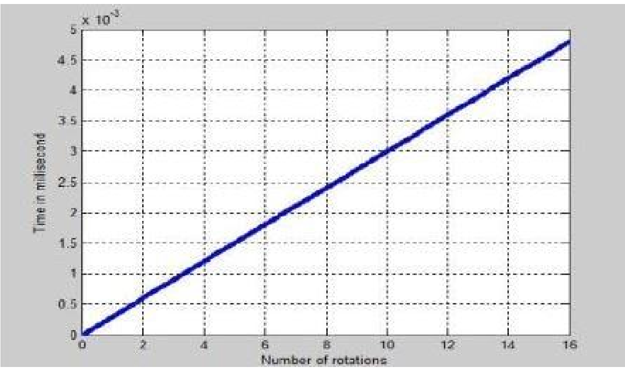


Fig. 13. Graphical representation of number of rotation vs. time (m sec.)

B. Time in numbers of rotations

Wadi and Zainal in [14] recently proposed an S-box based on modified AES-128 block cipher which is too easy to break which is claimed in [14]. So, the proposed 3-Dimensional S-box is dynamic, it depends on rotation. Rotation also takes some times, which has a great impact on computational time. Table III is showing the time in millisecond for rotations where rotation is possible from 0 to 16.In security analysis, it will take too long time for brute force approach. From the above table, it is observed that time is reasonable for any kind of rotation over S-box as it is dynamic. From the table III, we draw a graph on Fig. 13 in order to show the graphical representation of number of rotation vs. time in milliseconds.

TABLE III. COMPUTATIONAL TIME FOR ENCRYPTION IN SEC. VS FILE SIZE IN KB

Number of Rotation	Times(msec.)
0	0.000
1	0.003
2	0.006
3	0.009
4	0.012
5	0.015
6	0.018
7	0.021
8	0.24
9	0.027
10	0.030
11	0.033
12	0.036
13	0.039
14	0.042
15	0.045
16	0.048

C. Average time for different Galois Field (GF)

Time to encrypt and decrypt is an important feature of any encryption algorithm. As, S-box is a part of encryption algorithm, so, time has great impact on this dynamic 3-dimensional S-box. As we work in GF (3^5)), we want to show the time of other n=1,2,3,4,5 which is less than 6 because 3^6 creates some additional complexity which cannot be solved. Table IV is showing the average time of different GF$(3^n)$where n is an integer and is less than 6. From the table IV, we draw a graph on Fig. 14 in order to show the graphical representation of time vs. length of bits.

TABLE IV. TIME IN MILLISECOND OF GF(3^N) WHEN N<6

Length of bits	Average time to generate dynamic 3-dimensional S-box (ms)
3	0.0003
9	0.0057
81	0.0285
243	0.057

IX. FUTURE PERSPECTIVES AND CONCLUSION

Many attempts [15] have been made to modify AES algorithm based on GF(2^8). All of these modifications did not show the computational time of rounds and intermediate tasks on which the feasibility of AES depends. As it is known that key scheduling does not effect on time of encryption and decryption, we can make complex calculations. But in this paper, we mainly focus both on key scheduling as well as

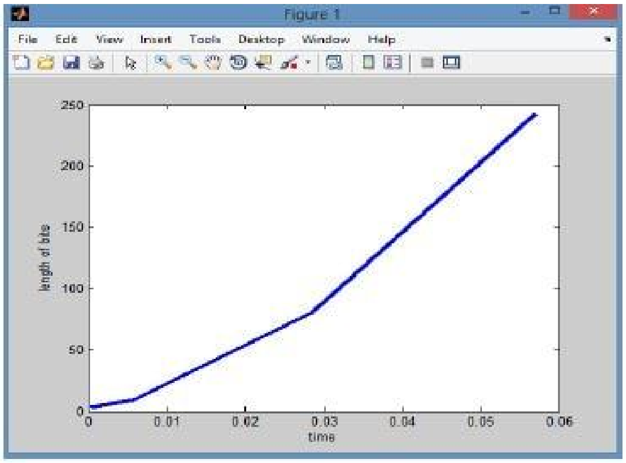


Fig. 14. Graphical representation of time (ms) vs. length of bits

S-box. A dynamic 3-dimensional S-box is created based on multiplicative polynomial inverse over GF(3^5) with 243 bits plaintext. In which, a new Substitution box (S-box) is proposed. For better security, we made it little complex keep pace with the computational time which is not so high. By thinking about security, we believe these dynamic 3-dimensional S-box and 3-dimensional cube key generation process can be used instead of traditional S-box. The limitation of this system is, it's concerned for the security not over the total required time. So this algorithm can be used for secure message transform, where time is not a conscious matter, like Governmental information transfer or Military data transfer. On the other hand, for fast transfer process, this system required high bandwidth.

In future, this system will be devolved to the image encryption standard based with 3-dimensional process. Up Next task of this system would be adding the authentication part for data security over cloud computing. At that stage the system will be concerned about the performance.

References

[1] S. Rajput, J. Dhobi, and L. J. Gadhavi, "Enhancing data security using aes encryption algorithm in cloud computing," in *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*. Springer, 2016, pp. 135–143.

[2] V. Rijmen and J. Daemen, "Advanced encryption standard," *Proceedings of Federal Information Processing Standards Publications, National Institute of Standards and Technology*, pp. 19–22, 2001.

[3] N.-F. Standard, "Announcing the advanced encryption standard (aes)," *Federal Information Processing Standards Publication*, vol. 197, pp. 1–51, 2001.

[4] *TECHNOLOGY; U.S. Selects a New Encryption Technique - The New York Times.*, 2017(accessed 26-Jan- 2017), http://www.nytimes.com/2000/10/03/business/technology-us-selects-a-new-encryption-technique.html.

[5] A. E. Standard, "Federal information processing standard (fips) publication 197," *National Bureau of Standards, US Department of Commerce, Washington, DC*, 2001.

[6] A. Agarwal and A. Agarwal, "The security risks associated with cloud computing," *International Journal of Computer Applications in Engineering Sciences*, vol. 1, pp. 257–259, 2011.

[7] P. Rewagad and Y. Pawar, "Use of digital signature with diffie hellman key exchange and aes encryption algorithm to enhance data security in cloud computing," in *Communication Systems and Network Technologies (CSNT), 2013 International Conference on*. IEEE, 2013, pp. 437–439.

[8] J.-W. Han, C.-S. Park, D.-H. Ryu, and E.-S. Kim, "Optical image encryption based on xor operations," *Optical Engineering*, vol. 38, no. 1, pp. 47–54, 1999.

[9] M. T. Tran, D. K. Bui, and A. D. Duong, "Gray s-box for advanced encryption standard," in *Computational Intelligence and Security, 2008. CIS'08. International Conference on*, vol. 1. IEEE, 2008, pp. 253–258.

[10] M. Venkatesh, M. Sumalatha, and C. SelvaKumar, "Improving public auditability, data possession in data storage security for cloud computing," in *Recent Trends In Information Technology (ICRTIT), 2012 International Conference on*. IEEE, 2012, pp. 463–467.

[11] N. Aleisa, "A comparison of the 3des and aes encryption standards," *International Journal of Security and Its Applications*, vol. 9, no. 7, pp. 241–246, 2015.

[12] R. Gharshi, "Suresha. enhancing security in cloud storage using ecc algorithm," *International Journal of Science and Research (IJSR), India Online ISSN*, pp. 2319–7064, 2013.

[13] W. Stallings, *Cryptography and Network Security: Principles and Practice. (3rd ed.)*. Prentice Hall, Upper Saddle River, New Jersey,, 2003.

[14] S. M. Wadi and N. Zainal, "High definition image encryption algorithm based on aes modification," *Wireless personal communications*, vol. 79, no. 2, pp. 811–829, 2014.

[15] D. J. Bernstein and P. Schwabe, "New aes software speed records," in *International Conference on Cryptology in India*. Springer, 2008, pp. 322–336.

Cyclic Redundancy Checking (CRC) Accelerator for Embedded Processor Datapaths

Abdul Rehman Buzdar*, Liguo Sun*, Rao Kashif†, Muhammad Waqar Azhar‡, Muhammad Imran Khan†§
*Department of Electronic Engineering and Information Science
†Micro/Nano Electronic System Integration R & D Center (MESIC)
University of Science and Technology of China (USTC), Hefei, China
‡Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
§Department of Electronics Engineering, University of Engineering and Technology Taxila, Pakistan

Abstract—**We present the integration of a multimode Cyclic Redundancy Checking (CRC) accelerator unit with an embedded processor datapath to enhance the processor performance in terms of execution time and energy efficiency. We investigate the performance of CRC accelerated embedded processor datapath in terms of execution time and energy efficiency. Our evaluation shows that the CRC accelerated Microblaze SoftCore embedded processor datapath is 153 times more cycle and energy efficient than a datapath lacking a CRC accelerator unit. This acceleration is achieved at the cost of some area overhead.**

***Keywords*—CRC; Accelerator; Codesign; FPGA; MicroBlaze; Embedded Processor**

I. Introduction

For reliable data communication Cyclic Redundancy Checking (CRC) is a well-known technique for error detection. The CRC calculations requires limited hardware resources and can be implemented easily. This is the reason that CRC is being used in industry for three decades, in spite the fact that more advance techniques for error detection and correction have been developed e.g. Viterbi decoder, low-density parity-check (LDPC), Reed Solomon and Turbo codes [1], [2], [3], [4].

Generally CRC can be implemented in software and executed on an embedded processor which requires a lot of clock cycles for the computation of CRC. The CRC can be implemented more efficiently in dedicated hardware which will require few clock cycles for the computation of CRC with some area overhead. The high speed communication systems today requires fast data rates which can only be delivered using dedicated hardware solutions.

Different hardware modules like USB, Ethernet, TCP/IP and CAN protocol are included in modern embedded processors to speedup certain parts of application in areas like signal processing, communication and control systems. All these protocols uses CRC for error detection. The addition of CRC accelerator into the embedded processor datapath will help to improve the overall performance. The commercially available off the shelf microcontrollers [5] and DSPs [6], [7] contain CRC hardware accelerator blocks.

II. CRC Computation Techniques

The computation of CRC is remainder of long modulo-2 division of input polynomial with a key polynomial. The CRC operation is performed in hardware using exclusive-OR and shift operations. The hardware implementation of CRC operation is composed of Exclusive-OR gates computational network which gives remainder and the registers for storage and shifting of current state, shown in Fig. 1. The key polynomial decides the width of state register e.g. for 16 bit key polynomial the width of state register will be 16 bits. The exclusive-OR gate network size depends on the input width and the technique used for the computation of CRC function.

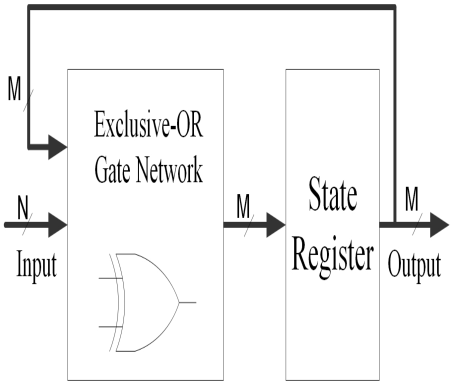


Figure 1: General architecture of a CRC computation circuit.

In serial implementation of CRC function Serial Linear Feedback Shift Registers (LFSR) are used which take only one input bit at a time. The serial implementation of CRC results into smaller exclusive-OR gate network but they are very slow as only one bit is shifted into LFSR circuitry in each cycle. Now a days parallel implementations of CRC are mostly used as they deliver faster speed. Fig. 2 shows the serial LFSR circuit implementation of key polynomial Equation (1). Every exponent of key polynomial is converted into an exclusive-OR gate between input and feedback path. This implementation is very slow as it accepts one bit at each cycle.

$$p(x) = x^5 + x^3 + x + 1 \tag{1}$$

The unfolding methodology [8], [9] can be used to implement parallel CRC circuitry which also uses LFSR as basic building block. Equation (1) can be converted into parallel CRC circuit using these unfolding techniques, as shown in Fig. 3. This parallel implementation of CRC gives twice speedup

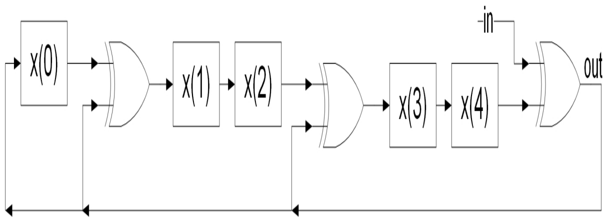


Figure 2: CRC circuit based on LFSR, implementing the key in Eq. 1.

compared to serial LFSR. This unfolding technique can be used to implement higher order parallelism which gives more speedup. But this technique has a drawback of higher fan-out as we increase the order of parallelism [10].

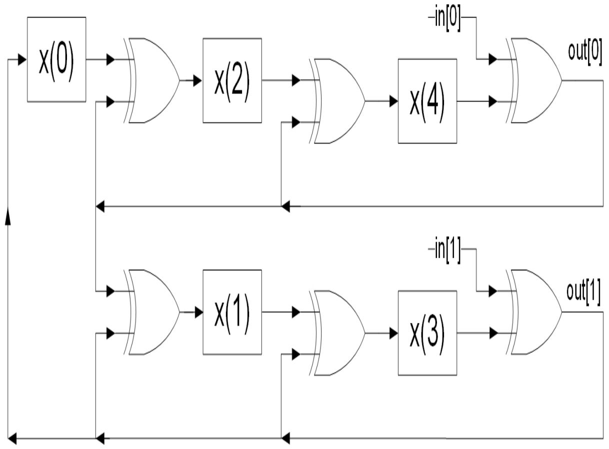


Figure 3: 2-level, unfolded CRC circuit that accepts two input bits each cycle.

A more efficient technique called state-space transformation can be used to implement parallel CRC circuits [11]. We implemented the parallel CRC circuits used in this work by following a technique invented by Campobello et al [12], shown in Fig. 4.

III. CRC Accelerator Implementation

We have implemented a 32-bit accelerator unit by including commonly used CRC circuits i.e. CRC5, CRC8, CRC16 and CRC32 inside a CRC accelerator main block. As we want to integrate this CRC accelerator unit with an embedded processor datapath, so it should be able to perform commonly used CRC operations. The required CRC operation can be selected using a 2-bit control signal. This configurable CRC accelerator unit is depicted in Fig. 5 and it can perform the following CRC operations [13], [14]:

1) **00**: CRC5 for USB interface.

$$p(x) = x^5 + x^2 + 1 \tag{2}$$

2) **01**: CRC8 for ATM protocols, etc.

$$p(x) = x^8 + x^2 + x + 1 \tag{3}$$

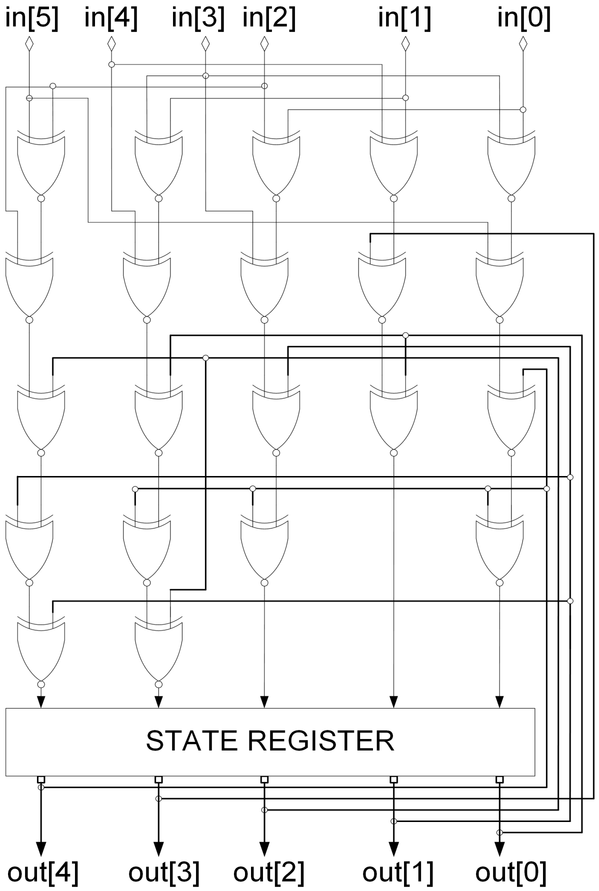


Figure 4: 6-bit input parallel CRC circuit for the key in Eq. 2.

3) **10**: CRC16 for XMODEM, X25 protocols, etc.

$$p(x) = x^{16} + x^{12} + x^5 + 1 \tag{4}$$

4) **11**: CRC32 for IEEE 802.3 standard.

$$\begin{aligned} p(x) =& x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + \\ & x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1 \end{aligned} \tag{5}$$

After performing the power analysis of CRC accelerator unit, shown in Fig. 5. We found that this design is not power efficient. The reason of this power inefficiency is the unnecessary switching taking place in all the CRC blocks. Because only one CRC sub-unit is required for the computation of desired CRC operation. So we decided to design a more power efficient CRC accelerator unit by disabling the CRC blocks which are not in use by employing power gating technique. We also used distributed multiplexer at the output of CRC accelerator unit and clock gating for disabling the registers which are not in use to save the switching power. The power efficient CRC accelerator unit is shown in Fig. 6. Both the CRC Accelerator units were verified and synthesized using Xilinx ISE Design Suit [15]. The Initial CRC unit and Low Power CRC hardware accelerator block were synthesized on 7vx485tffg1157-3 Virtex-7 FPGA device which is based on a 28nm technology. It gives a critical path delay of 1.868ns and 1.986ns respectively. The synthesis results are shown in

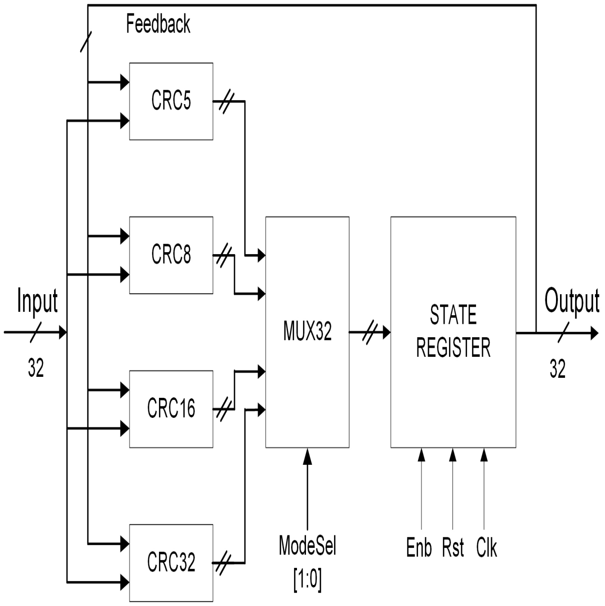


Figure 5: First CRC accelerator unit.

Table I. As can be seen the low power CRC accelerator is power efficient compared to the initial CRC unit.

Table I: Synthesis Results of CRC Accelerator Units

	First CRC Unit	Low Power CRC Unit
Power	523mW	241mW
Max Freq	535.217MHz	503.499MHz
Latency	1.868ns	1.986ns
Slice Registers	32	32
Slice LUTs	229	340
Occupied Slices	98	166

IV. Integration of CRC Accelerator Unit with MicroBlaze Processor

We have implemented the CRC accelerator unit in VHDL hardware description language and verified it using Xilinx ISE design suit [15]. We used Xilinx Spartan-6 FPGA SP605 Evaluation Kit [17] and Xilinx Embedded Development Kit (EDK) [15] for the implementation. The Hardware/Software co-design is a well established technique which improves the performance of the system [16-19]. Xilinx Microblaze soft core processor [16] was used to run the software implementation of CRC. There are two ways to integrate a hardware accelerator core into a MicroBlaze-based embedded soft processor system. One way is to connect the accelerator through the Processor Local Bus (PLB). The second way is to connect it using MicroBlaze dedicated Fast Simplex Link (FSL) bus system [18]. First PLB was tried but it was taking a lot of cycles. Because it is a traditional memory mapped transaction bus. Then it was decided to integrate our CRC accelerator unit using a dedicated FIFO style FSL Bus with the MicroBlaze processor system, shown in Fig. 7.

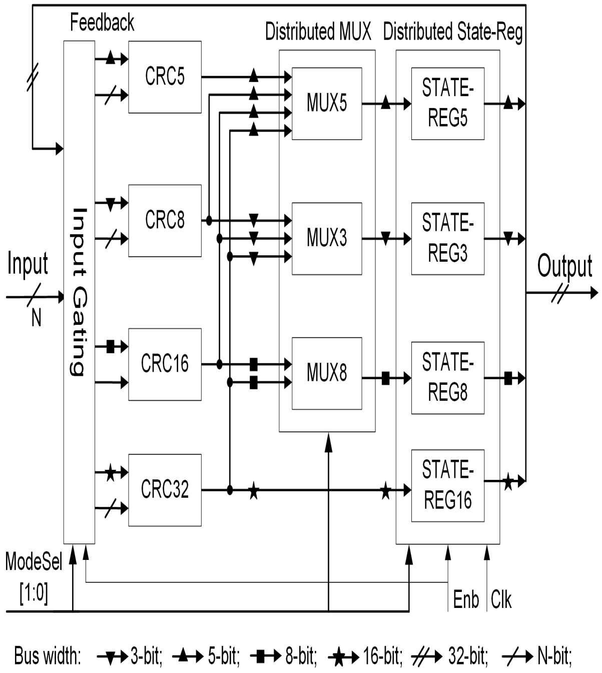


Figure 6: Low-power CRC accelerator unit.

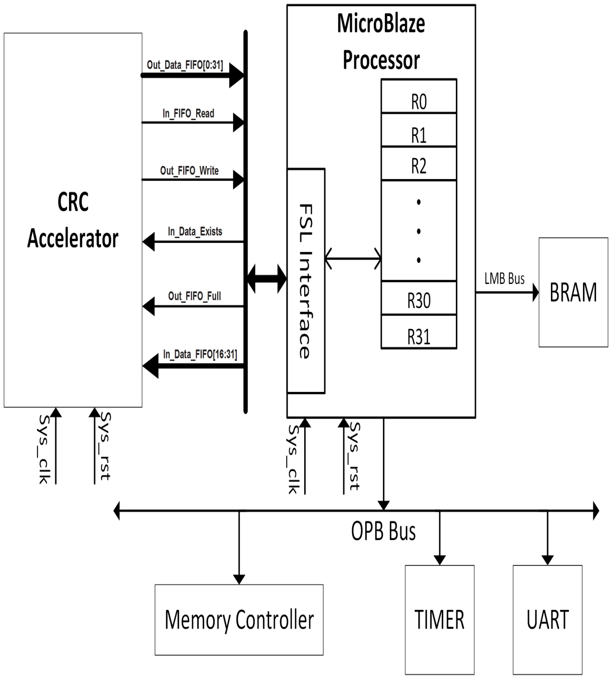


Figure 7: CRC Accelerator Unit with MicroBlaze Processor System

The software only C code for CRC5, CRC8, CRC16, CRC32 was implemented and verified. Later these C codes were executed on the MicroBlaze processor using Xilinx Software Development Kit (SDK) [15]. The cycle count for the complete software implementations of CRC was measured using the XPS hardware timer block, shown in Table II. Fig. 8 and 9 shows the cycle count and energy dissipation of different architectures, respectively.

Table II: Cycle Count and Energy Dissipation at Clock Period 20ns

Architecture	#Cycles	Power (mW)	Energy* (μJ)
CRC5 SW	1086	178	3.8661
CRC8 SW	2652	178	9.4411
CRC16 SW	5200	178	18.512
CRC32 SW	5373	178	19.1278
CRC HW	35	184	0.13

*: Energy = #cycles × clock period × power.

The CRC accelerator unit was attached with the Microblaze processor system via FSL bus using Xilinx Platform Studio (XPS) [15]. The software part of CRC accelerator unit was implemented in C programming with Xilinx SDK. The predefined C functions of SDK were used to communicate with hardware part of CRC accelerator unit via FSL bus. Our evaluation shows that an accelerated MicroBlaze processor datapath is 153 times more cycle and energy efficient than a datapath lacking CRC accelerator.

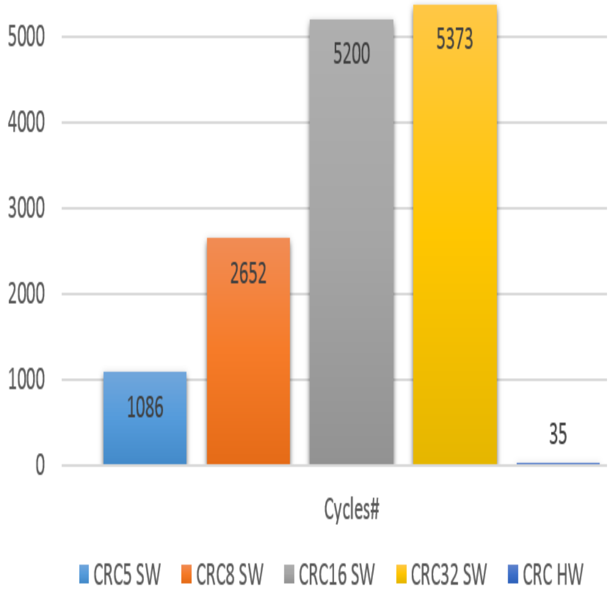


Figure 8: Cycle count of various CRC implementations.

V. Conclusion

In this paper, we have designed a flexible CRC accelerator unit using VHDL. We have integrated the CRC accelerator unit with the Microblaze Softcore processor system using FSL Bus to enhance the processor performance. We used Xilinx Spartan-6 FPGA Evaluation Kit and Xilinx Embedded Development Kit (EDK) for the implementation. We have shown that a CRC accelerated Microblaze embedded processor datapath is 153 times more cycle and energy efficient that a datapath lacking a CRC accelerator with some area overhead.

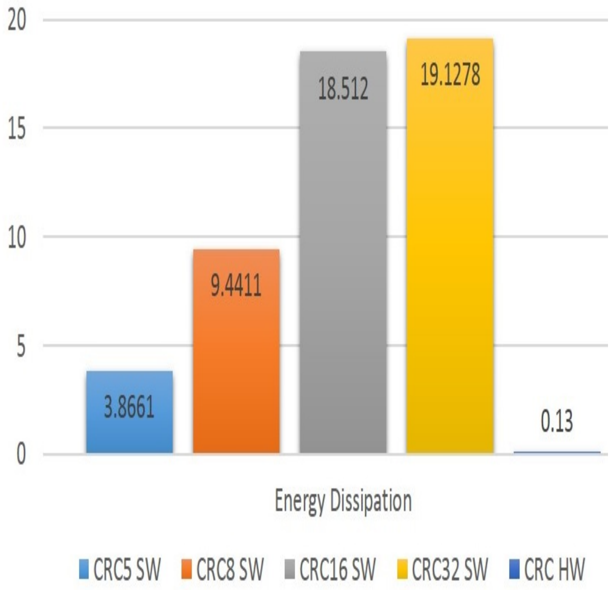


Figure 9: Energy dissipation of various CRC implementations.

Acknowledgment

This work is partially supported by the Chinese Academy of Sciences and The World Academy of Sciences CAS-TWAS President's Fellowship 2013-2017.

References

[1] M. F. Brejza, L. Li, R. G. Maunder, B. Al-Hashimi, C. Berrou, L. Hanzo, "20 years of turbo coding and energy-aware design guidelines for energy-constrained wireless applications", IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 8-28, 1st Quart. 2016.

[2] Mehran Mozaffari Kermani, Vineeta Singh, Reza Azarderakhsh, "Reliable Low-Latency Viterbi Algorithm Architectures Benchmarked on ASIC and FPGA," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 64, no. 1, pp. 208-216, 2017.

[3] Linjia Chang, Avhishek Chatterjee, Lav R. Varshney, "Performance of LDPC Decoders With Missing Connections," IEEE Transactions on Communications, vol. 65, no. 2, pp. 511-524, 2017.

[4] Salvatore Pontarelli, Pedro Reviriego, Marco Ottavi, Juan Antonio Maestro, "Low Delay Single Symbol Error Correction Codes Based on Reed Solomon Codes," IEEE Transactions on Computers, vol. 64, no. 5, pp. 1497-1501, 2015.

[5] Atmel, "Secure microcontroller for smart cards." [Online]. Available: http://www.atmel.com

[6] Freescale, "MAPLE hardware accelerator and SC3850 DSP core." [Online]. Available: http://www.freescale.com

[7] Microchip, "PIC32mx775f512l datasheet." [Online]. Available: http://www.microchip.com

[8] K. K. Parhi, VLSI Digital Signal Processing Systems - Design and Implementation. Wiley-Interscience Publishers Inc., 1999.

[9] C. Cheng and K. K. Parhi, "High-Speed Parallel CRC Implementation Based on Unfolding, Pipelining, and Retiming," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 53, no. 10, pp. 1017-1021, 2006.

[10] T.-B. Pei and C. Zukowski, "High-Speed Parallel CRC Circuits in VLSI," IEEE Transactions on Communications, vol. 40, no. 4, pp. 653-657, Apr. 1992.

[11] J. H. Derby, "High-Speed CRC Computation Using State-Space Transformations," in IEEE International Global Telecommunications Conference, Nov. 2001, pp. 166-170.

[12] G. Campobello, G. Patane, and M. Russo, "Parallel CRC Realization," IEEE Transactions on Computers, vol. 52, no. 10, pp. 1312-1319, Oct. 2003.

[13] E. Stavinov, "A practical parallel CRC generation method." [Online]. Available: www.OutputLogic.com

[14] Muhammad Waqar Azhar, Tung Thanh Hoang, and Per Larsson-Edefors, "Cyclic Redundancy Checking (CRC) Accelerator for the Flex-Core Processor," in Proc. of EUROMICRO Conf. on Digital System Design, 2010, pp. 675-680.

[15] Xilinx Inc. FPGA Design Tools. Silicon Devices. [Online]. Available: http://www.xilinx.com

[16] Xilinx MicroBlaze [Online] www.xilinx.com/tools/microblaze.htm

[17] Xilinx Spartan-6 FPGA SP605 Evaluation Kit. [Online] Available: www.xilinx.com/products/boards-and-kits/ek-s6-sp605-g.html

[18] Xilinx Fast Simplex Link (FSL). [Online] Available: http://www.xilinx.com/products/intellectual-property/fsl.html

[19] Abdul Rehman Buzdar, Liguo Sun, Azhar Latif and Abdullah Buzdar, "Distance and Speed Measurements using FPGA and ASIC on a high data rate system" International Journal of Advanced Computer Science and Applications(IJACSA), 6(10), 2015, pp.273-282.

[20] Abdul Rehman Buzdar, Liguo Sun, Azhar Latif and Abdullah Buzdar, "Instruction Decompressor Design for a VLIW Processor", Informacije MIDEM-Journal of Microelectronics, Electronic Components and Materials Vol. 45, No. 4 (2015), pp.225-236.

[21] Abdul Rehman Buzdar, Azhar Latif, Liguo Sun and Abdullah Buzdar, "FPGA Prototype Implementation of Digital Hearing Aid from Software to Complete Hardware Design" International Journal of Advanced Computer Science and Applications(IJACSA), 7(1), 2016, pp.649-658.

[22] Abdul Rehman Buzdar, Liguo Sun, Shoab Ahmed Khan, Abdullah Buzdar, "Area and Energy efficient CORDIC Accelerator for Embedded Processor Datapaths" Informacije MIDEM-Journal of Microelectronics, Electronic Components and Materials Vol. 46, No. 4(2016), pp.197-208

Prediction by a Hybrid of Wavelet Transform and Long-Short-Term-Memory Neural Network

Putu Sugiartawan, Reza Pulungan, and Anny Kartika Sari
Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences
Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—Data originating from some specific fields, for instance tourist arrivals, may exhibit a high degree of fluctuations as well as non-linear characteristics due to time varying behaviors. This paper proposes a new hybrid method to perform prediction for such data. The proposed hybrid model of wavelet transform and long-short-term memory (LSTM) recurrent neural network (RNN) is able to capture non-linear attributes in tourist arrival time series. Firstly, data is decomposed into constitutive series through wavelet transform. The decomposition is expressed as a function of a combination of wavelet coefficients, which have different levels of resolution. Then, LSTM neural network is used to train and simulate the value at each level to find the bias vectors and weighting coefficients for the prediction value. A sliding windows model is employed to capture the time series nature of the data. An evaluation is conducted to compare the proposed model with other RNN algorithms, *i.e.*, Elman RNN and Jordan RNN, as well as the combination of wavelet transform with each of them. The result shows that the proposed model has better performance in terms of training time than the original LSTM RNN, while the accuracy is better than the hybrid of wavelet-Elman and the hybrid of wavelet-Jordan.**

Keywords*—*Wavelet Transform; Long-Short-Term Memory; Recurrent Neural Network; Time Series Prediction

I. Introduction

The growth in the number of visitors and tourism investments makes tourism become a key factor in export earnings, job creation, business development and infrastructure. Tourism has shifted and become one of the largest fast growing economic sectors in the world. Despite the global crises that occur several times, the number of international tourist trips continues to show positive growth. As shown by the data from BPS, the Indonesian Central Agency for Statistics, the number of tourists visiting Indonesia has increased from year to year. Travel and tourism directly contributes 2.1 trillion dollars to global GDP. It is more than doubled, compared to the automotive industry, and nearly 40 percent larger than the global chemical industry [1]. Travel and tourism sector is worth three quarters of the education sector, the banking sector, the mining sector, and the communications services sector. By knowing the number of the visitors to a country, the income of the country from the tourism sector can be predicted.

Fluctuation in the number of tourists visiting Indonesia in every year is not easy to predict. This has become a major problem for some parties such as hotels, restaurants and travel agents. This also causes those parties not able to devise good plannings for their business. The difficulty in determining the data traffic patterns is due to the existence of noise.

To overcome this problem, a technique is needed to separate the low frequency pattern from the high frequency pattern through the process of translation (shifting) and dilation (scaling) [2]. Wavelet transform can reveal aspects of frequencies in the frequency decomposition process [2], [3]. In [4], the use of wavelet sequence prediction models improves the effectiveness of multi layer perceptron (MLP) neural network. A merger between wavelet model and Kalman filter produces a powerful model for estimation technique [5], so does the merging of wavelet model with spectral analysis [6]. Evaluations conducted to several wavelet-RNN models show that the combination between wavelet and RNN usually produces smaller error value [7].

Recurrent neural networks have the capability to dynamically incorporate past experience due to internal recurrence [2]. RNNs can project the dynamic properties of the system automatically, so they are computationally more powerful than feed-forward networks, and the valuable approximation results are obtained for chaotic time series prediction [8], [9]. One of RNN models is long-short-term memory. This model works when there is a long delay, and is able to handle signals that have a mixture of low and high frequency components. The learning process of RNN models however requires a relatively long time because there is a context layer in the network architecture [10].

LSTM is a successful RNN architecture model to fix the vanishing gradient problem in neural network [11]. Sequence-based prediction of protein localization produces high prediction accuracy with LSTM and bidirectional model [12]. Comparing LSTM RNN model with random walk (RW), support-vector machine (SVM), single-layer feed forward (FFNN) and stacked autoencoder (SAE) shows that the LSTM RNN model produces higher accuracy and generalizes well [13]. Prediction of time series data for securities in Shanghai *ETF180* obtains a good accuracy. Using LSTM RNN, the result increases by 4 percents compared to the previous model, while data normalization can also improve accuracy [14].

In this paper a new hybrid algorithm for prediction, which is based on a combination of wavelet analysis and LSTM neural networks, is proposed. The proposed method is then applied for predicting tourist arrivals. The wavelet is employed to denoise the original signal and decompose the historical number of tourist arrivals into better series pattern for prediction. An LSTM neural network is used as a non-linear pattern recognition to estimate the training data signal and to compensate the error of wavelet-LSTM prediction. The proposed method is applied to tourist arrival data, which is a set of 240 vector data.

The rest of the paper is organized as follows: The principles of the proposed method are described in Section II. Simulation result and the comparison of the proposed model with Elman and Jordan recurrent neural networks are presented in Section III. Finally, we conclude the paper and present future work in Section IV.

II. Methodology

Improving the accuracy of prediction can be performed by combining several different methods [15]. In this paper, LSTM neural network model is used to identify data pattern, while the wavelet method is employed to decompose input data. Prediction model using the hybrid of wavelet transform and LSTM neural network consists of the following phases:

- Phase 1: normalizing the data to values ranging between 0 and 1,
- Phase 2: decomposing data into constitutive series through wavelet transform,
- Phase 3: applying sliding windows to the data to form several variables, and
- Phase 4: recognizing data pattern using LSTM neural network model through data training and data testing.

The result of the proposed model is then compared with Elman RNN, Jordan RNN and LSTM RNN, as well as the hybrid of wavelet-Elman and the hybrid of wavelet-Jordan. The following subsections further elaborate the general design of the proposed model, the details of the wavelet transform as well as LSTM, data normalization and the evaluation metrics used.

A. Design of the proposed model

Fig. 1 depicts the flowchart of the proposed model. The model is divided into two processes, *i.e.*, training and testing, each of which further contains several processes. The first process in data training is to normalize the data with Min-Max normalization. The second process is data decomposition using wavelet transform. The purpose of wavelet transform is to divide the data into high and low frequencies. The third process is segmentation of time series data input using sliding windows. The next process is training using LSTM RNN. The training process uses 90 percent of data. The result of training process is a weight value for each neuron in the neural network and error value, which are shown by MSE and RMSE. After the weight value for each neuron is obtained, the testing process is started. The testing process produces error value. To restore the decomposed values to the original values, reconstruction and denormalization processes are performed. The purpose of reconstruction is to restore the data from high and low frequencies, while denormalization aims at restoring the data to original values.

In neural networks, data training is performed to gain bias values and weightings for prediction approximation and detail coefficients [3]. Bias and weighting coefficients generated from learning data are used in the testing process. The process is conducted iteratively to generate the prediction coefficient details. Another learning process with the purpose of obtaining a prediction coefficient approximation has the same initial step. The learning does not use the input from the beginning, but is performed on approximation coefficients at the lowest levels (DWT 3). Once the simulation result has reached the desired target or maximum epoch, the learning finishes. The learning process uses back-propagation algorithm with the addition of a context layer [16] to accelerate the convergence towards the desired minimum value error.

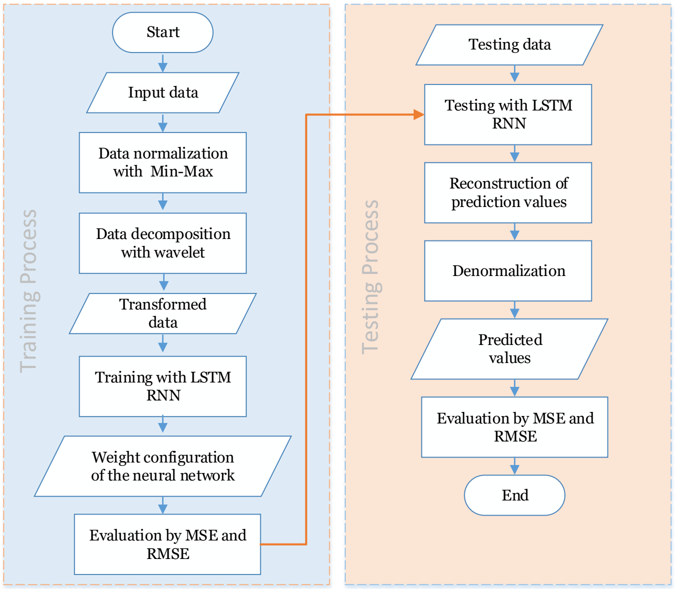

Training Process
Start
Input data
Data normalization with Min-Max
Data decomposition with wavelet
Transformed data
Training with LSTM RNN
Weight configuration of the neural network
Evaluation by MSE and RMSE
Testing Process
Testing data
Testing with LSTM RNN
Reconstruction of prediction values
Denormalization
Predicted values
Evaluation by MSE and RMSE
End

Fig. 1. Flowchart of the proposed model

B. Data normalization

Prior to the decomposition process, research data obtained from BPS must be normalized. The technique used provides linear transformation on original range of data, and is called Min-Mix normalization [17], [18]. The technique keeps the relationship among original data. Min-Max normalization is a simple technique, where the technique can specifically fit the data in a pre-defined boundary. A normalized value $\hat{x}$ of a data point x_i in a predefined boundary $[C, D]$ is defined by:

$$\hat{x} = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}(D - C) + C, \tag{1}$$

where x_{min} is the smallest value of the data, x_{max} is biggest value of the data, and $[x_{min}, x_{max}]$ is the range of the original data. The normalization process will produce a value ranging between 0 and 1. The biggest value will produce a value of 1, while the smallest value will produce a value of 0 in the normalization process.

To restore the normalization value to the original value, denormalization process is conducted. This aims at restoring the output of the value to be in the range beforehand. Given a normalized value $\hat{x}$, its denormalization, namely, the original data point x_i can be calculated by:

$$x_i = \frac{\hat{x}(x_{max} - x_{min})}{(D - C) + C} + x_{min}. \tag{2}$$

C. Data decomposition with wavelet transform

A wave is a function that moves up and down space at a time on a periodic basis, while wavelet is a restricted or localized wave [19]. Wavelet can also be regarded as a short-wave. The model provides a depiction of the frequency of the signal timing.

A wavelet transform (WT) is a time-frequency decomposition that provides a useful basis of time series in both time and frequency [8], [20], when the time series, like tourist arrival series, is non-stationary. Mother wavelet is the basic function used in wavelet transform [2] as it produces all functions used in the wavelet transformation through translation and scaling. The mother wavelet will determine the characteristics of the produced wavelet transform. Therefore, selection of the mother wavelet type must be done carefully in order to perform the transformation efficiently.

The wavelet transform can identify and analyze the signal moves. The purpose of analyzing the signal moves is to obtain information and the frequency spectrum at the same time. Discrete wavelet transform (DWT) is one of the wavelet transform development series. DWT works on two collections of functions called scaling functions and wavelet functions that are each associated with a low-pass filter and a high-pass filter [20]. The decomposition structure of wavelet transforms for level 3 is shown in Fig. 2.

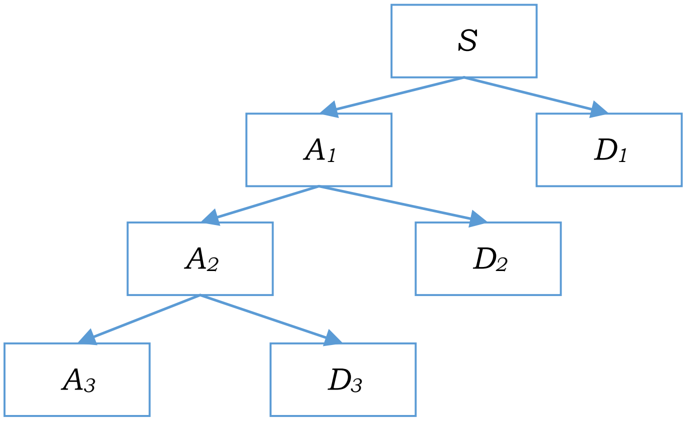


Fig. 2. Wavelet decomposition tree at level 3

The type of wavelet used in this research is *Haar* wavelet transform. It is a simple Daubechies wavelet, which is suitable to detect time-localized information and increases the performance of the prediction technique [21]. The Haar wavelet has two functions called approximation function and difference function. The approximation function produces a sequence of the averages between two consecutive data in the data input, while the difference function produces the current approximation sequence. Both functions are executed recursively and the process will stop when the element in the difference sequence is one [22]. The i-th approximation sequence (A_i) is given by:

$$A_i = \frac{A_{i-1}(1) + A_{i-1}(2)}{2} + \frac{A_{i-1}(3) + A_{i-1}(4)}{2} + \dots + \frac{A_{i-1}(n-1) + A_{i-1}(n)}{2}, \quad (3)$$

where $A_{i-1}(j)$ is the j-th element in the sequence (A_{i-1}) for $j = 1, 2, \dots, n$.

Decomposition process can be through one or more levels. Discrete wavelet series contain approximated series (A_t) and detail series (D_t). Dimensional signal can be divided into two parts, the high frequency and the low frequency parts. The high frequencies is analyzed with low-pass filter, while the low frequencies is analyzed with high-pass filter. Both frequency filters are used to analyze the different resolutions of the signal. The signal could be subsampled by 2, by discarding every second sample. The decomposition for each layers are represented by [8]:

$$y_{low}(k) = \sum_i x(i)h(2k-i), \text{ and} \quad (4)$$

$$y_{high}(k) = \sum_i x(i)g(2k-i), \quad (5)$$

where y_{low} and y_{high} is a low-pass and high-pass filters, respectively, both subsampling by 2 [8]. k refers to the time decomposition and the original signal data $x(i)$ is passed through to a high-pass filter $g(\cdot)$ and a low-pass filter $h(\cdot)$. The above functions can be reused in the next decomposition. DWT coefficient consists of the output of high-pass and low-pass filters.

The high and low-pass filter functions are followed in reverse order by the reconstruction. The signal at each layer is upsampled by two, through the synthesis filter $g'(\cdot)$ and $h'(\cdot)$ and added to each other [8]. The reconstruction for each layer is given by:

$$x(i) = \sum_k y_{high}(k)g'(-n+2k) + y_{low}(k)h'(-n+2k). \quad (6)$$

The reconstruction process aims at returning the original values of data. Reconstruction is started by combining DWT coefficients which are at the end of the previous decomposition upsampled by 2 ($\uparrow 2$) through a high-pass filter and low-pass filter. The reconstruction process is completely the opposite of the decomposition process according to the level of decomposition [19].

D. Sliding windows technique

Sliding windows technique is a kind of processing method of concept drift in data streams, which has many applications in intrusion detection. The essence of this technique is data update mechanism. The data stream (x) is divided into several parts of data blocks. When sliding window moves to the next block, new block is added to the window at intervals, and the oldest block is deleted. Through this dynamic sample selection method, the sample for modeling is updated [23].

The technique of sliding windows comes with a particular size of window; and this impacts the size of sample data. Suppose that $\langle x_0, x_1, x_2, \dots, x_{n-1}, x_n, x_{n+1}, \dots \rangle$ is a series of time-series data. When the window size is fixed at k, the data interval will be changed to $\langle x_{i-k}, x_{i-k+1}, \dots, x_i, x_{i+1} \rangle$ and has different data streams from older data streams. As shown in Fig. 3, the data interval is $\langle x_{i-2}, x_{i-1}, x_i, x_{i+1} \rangle$ for window size of 3, where the value of x_{i+1} is obtained from $\{x_{i-2}, x_{i-1}, x_i\}$ values.

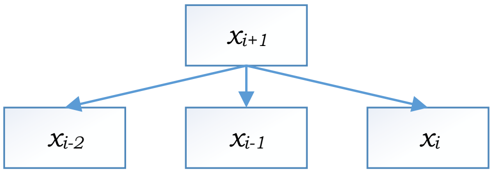


Fig. 3. Sliding windows technique with window size of 3

In the proposed model, the window size in the sliding windows depends on the number of data. The window size determines prediction accuracy. Based on several experiments, the window size selected in this paper is 3, because it provides the best value and minimizes data reduction. The number of data is 120 after decomposition process with wavelet. Then, the sliding windows process generates 3 data input variables, 1 output variable, and there will be 117 data. The number of data is reduced because the windows eliminates the last data record.

The implementation of sliding windows technique on RNN is shown in Fig. 4. The number of previous values used as input $\{\dots, x_{n-2}, x_{n-1}\}$ depends on the window size k, while the current value becomes the target y value at the output layer. Since the window size is 3, $\{x_{n-2}, x_{n-1}, x_n\}$ are the input values, while x_{n+1} is the output value of the RNN architecture.

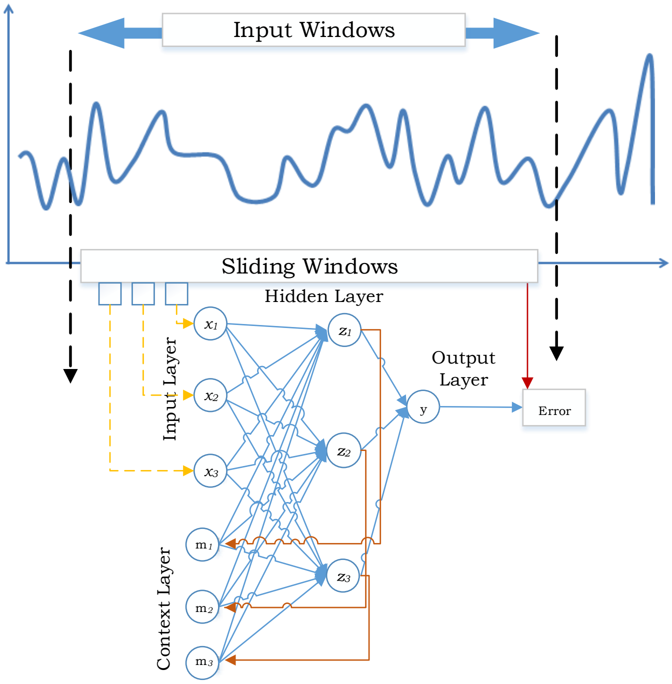


Fig. 4. Illustration of sliding windows technique on RNN

E. Data training with LSTM RNN model

Fig. 5 depicts the general architecture of an LSTM RNN [9]. An LSTM RNN is an artificial neural network structure for the solution of vanishing gradient. The algorithm works when there is a long delay, and can handle signals that have a mixture of low and high frequency components. The LSTM contains special units called memory blocks in the recurrent hidden layer as shown in Fig. 5. The memory blocks contain memory cells with self-connections, storing the temporal state of the network in addition to special multiplicative units called gates to control the flow of information.

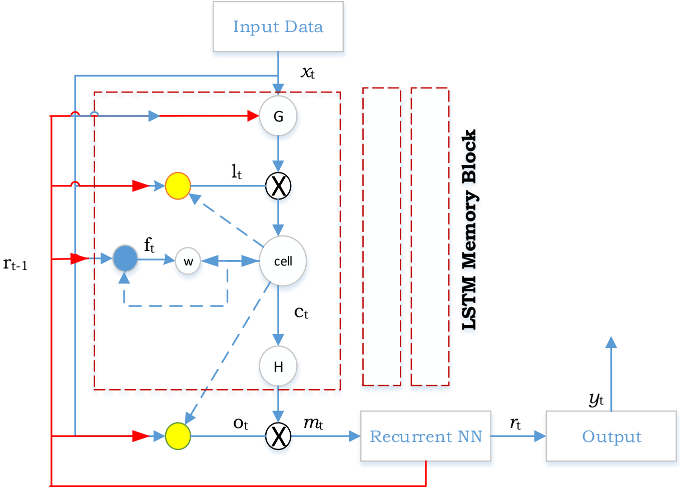


Fig. 5. The general LSTM-RNN architecture

Each memory block in the original architecture contains an input gate (G) and an output gate (H). The input gate (G) controls the flow of input activation (X_t) into the memory cell. The output flow (C_i) of cell activation is controlled by the output gate into the rest of the network; the next process is memory block added by forget gate (w) [24].

The forget gate (w) scales the internal state of the cell before adding it as input to the cell through the self-recurrent connection of the cell, therefore adaptively forgetting or resetting the cell's memory. In addition, modern LSTM architectures contain peephole connections from its internal cells to the gates in the same cell to learn the precise timing of the outputs [25].

An LSTM network computes from an input sequence $x = \langle x_1, \dots, x_T \rangle$ an output sequence $y = \langle y_1, \dots, y_T \rangle$ by calculating the network unit activations using the following equations, for $t = 1, \dots, T$ [25]:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i), \quad (7)$$
$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f), \quad (8)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c), \quad (9)$$
$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o), \quad (10)$$
$$m_t = o_t \odot h(c_t), \quad (11)$$
$$y_t = \emptyset(W_{ym}m_t + b_y). \quad (12)$$

We use W to denote the weight matrices (*e.g.*, W_{ix} is the matrix of weights from the input gate to the input), W_{ic}, W_{fc}, and W_{oc} are diagonal weight matrices for peephole connections. Vector b denotes a bias vector and b_i is the input gate bias vector. Function σ is the logistic Sigmoid function, i is the input gate, f is the forget gate, o is the output gate, and c denotes the cell activation vector. Functions i, f, o and c have the same size as the cell output activation vector m. Operator $\odot$ is the element-wise product of the vectors. Function g is the cell input function, while function h is a cell output activation function. Activation functions used in this paper are $\tanh$ and $\emptyset$, which denotes the network output activation function.

Fig. 6 shows the architecture of the LSTM-RNN in the proposed model. The number of hidden layer in the model is 1, hence the context layer must be 1. The training process is performed iteratively until the output value approximates the original value. The number of epoch in the training process is 10^5 with the learning rate of 0.1. $\{x_1, x_2, x_3\}$ is an input data at input layer, $\{z_1, z_2, z_3\}$ is a hidden layer unit, $\{m_1, m_2, m_3\}$ is a context layer unit, and y is an output value (the result of the RNN architecture).

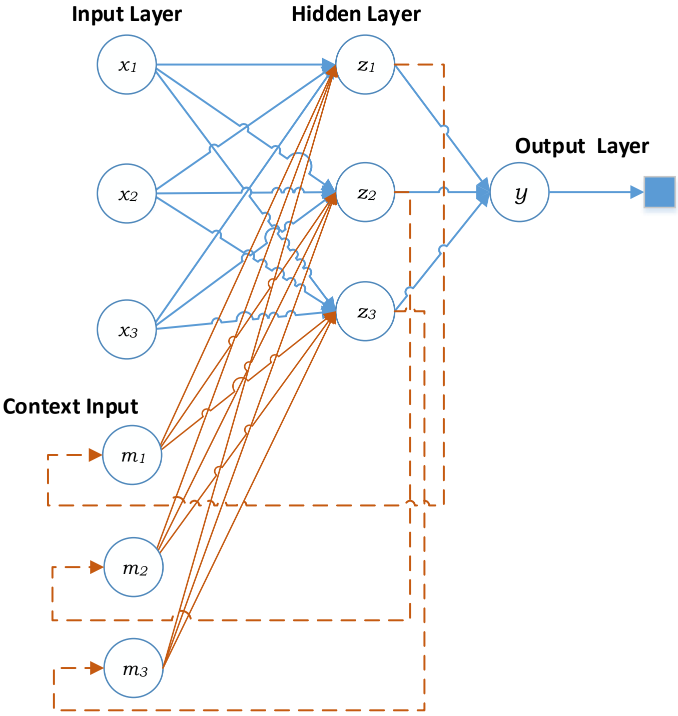


Fig. 6. The proposed LSTM-RNN architecture

F. Prediction accuracy

To measure and evaluate the prediction accuracy of the proposed hybrid model, mean square error (MSE) and root mean square error (RMSE) methods are used. Let y_i be the measured value of time i or the targeted minimum error on a neural network, f_i is the predicted value at time i obtained from a particular model M, and n be the number of sample data. Then, MSE is given by [26]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i), \tag{13}$$

and RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)}. \tag{14}$$

RMSE is used because it has a good performance to distribution error and could satisfy the triangle inequality requirement for a distance metric [27]. MSE is an error function for evaluating the performance and efficiency of the forecasting methods. MSE can compare point-by-point for overall performance measure method of the actual time series values and the forecast value [28]. If the values of RMSE and MSE are lower, the accuracy of the model is better.

III. Result and Analysis

The proposed model is applied to predict tourist arrivals in Indonesia. Data used in this research consists of the number of tourist visits to Indonesia from 1995 to 2014 in each month. Hence, there is 240 data records in total. Ninety percent of the data is for training, while 10 percent is for testing. The model is then compared to the original LSTM RNN and other types of RNN, *i.e.*, Elman RNN and Jordan RNN, as well as the hybrid of wavelet-Elman and the hybrid of wavelet-Jordan. These models will be compared in terms of accuracy and the time required for data training and data testing.

The first process is data normalization which aims to simplify calculations, reduce the value range, and make the training process faster. The normalization, which utilizes Min-Max technique, produces data in the range of $[0, \ldots, 1]$. Data is then transformed using Haar wavelet function, and 3 levels of decomposition is applied. The wavelet transform is implemented in Matlab.

Fig. 7 shows three levels of wavelet decomposition of tourist arrival time-series data. Graph S shows the time series data after normalization by Min-Max technique. The first level of decomposition process using Haar function produces high and low frequencies, as shown in graph A_1 and D_1. Data in each decomposition is divided into two, which each at level 3 produces 30 data records. The decomposition with level three A-series data (graph A_3) has the lowest frequency content and this tends to be incompatible with prediction.

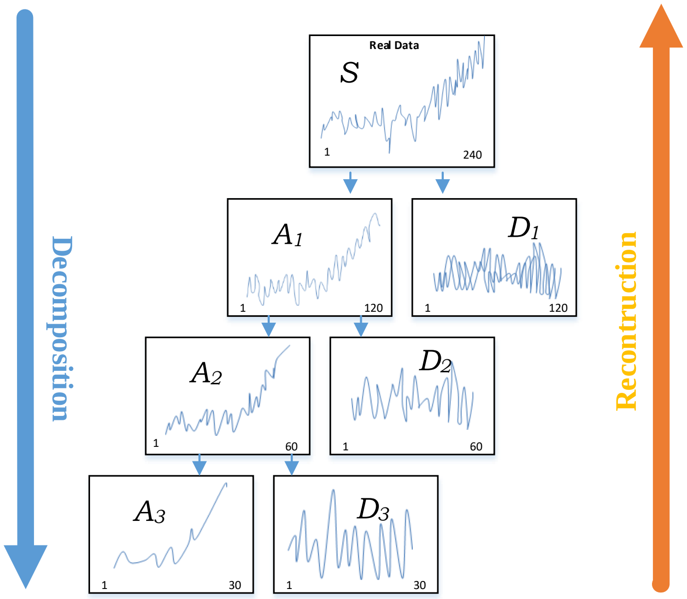


Fig. 7. Three-level wavelet decomposition of time series

The prediction is performed to data in each of graphs S, A_1, A_2 and A_3. Here, LSTM neural network with sliding windows technique with window size 3 is employed. The LSTM model includes 3 input data, a hidden layer with 4 LSTM blocks or neurons and an output layer that produces a single value prediction. The activation function for LSTM blocks is binary Sigmoid. The network is trained for 105 epochs and a batch size of 1 is used. Table I depicts LSTM parameters for data training. The prediction is implemented using Theano library provided in Python.

TABLE I. LSTM Parameters for Data Training

Parameter	Value	Information
Error Target (MSE)	0.001	Prediction error value
Learning Rate	0.01	Pace of learning
Hidden Layer	1	Number of hidden layers
Input Layer	3	Number of input layers
Output Layer	1	Number of output layers
Epoch Max	10^5	Maximum epoch
Transfer Function	Binary Sigmoid	
Weight	Random $(0, 1)$	

The proposed model is compared to hybrid wavelet-Elman, hybrid wavelet-Jordan, the original Elman RNN, Jordan RNN, and LSTM RNN. Tabel II depicts the results of the comparison between the proposed model and other models. The comparison between the models is based on the accuracy and the time required for data training and data testing. The result shows that the hybrid of wavelet transform and LSTM has better performance than other hybrid models in terms of accuracy. Nevertheless, the proposed model takes the most time for training process compared to other hybrid models. On the other hand, the hybrid model is able to reduce training time on the original LSTM model by 28 minutes. All hybrid models are able to reduce training time significantly; an average of twice faster than the training time required without the use of wavelet transform. Hybrid wavelet-Jordan produces the smallest training time among the models, but the error value is quite high compared to the original LSTM and the hybrid wavelet-LSTM model.

TABLE II. Comparison of Performance using Various Prediction Methods

Algorithm	MSE	RMSE	Training Time
Hybrid Wavelet and LSTM (proposed)	0.11853	0.34428	00:31:58
Hybrid Wavelet and Elman	0.14521	0.38107	00:12:37
Hybrid Wavelet and Jordan	0.21776	0.46664	00:09:52
LSTM RNN	0.00464	0.21541	00:59:11
Elman RNN	0.24791	0.49796	00:20:16
Jordan RNN	0.35256	0.59377	00:17:31

Fig. 8 depicts the comparison of the real data with the decomposition of signal produced using wavelet transform as well as the prediction of signal using LSTM neural network. In prediction where all data is used, the result is very similar to the original data. Due to decomposition, data is then reduced in each level of decomposition. Using data produced by DWT level 1, prediction result is still recognizable, and the accuracy is not too bad. However, using data of DWT levels 2 and 3, prediction results become more unrecognizable; hence, there is a big gap between real data and predicted data. This means that the deeper the decomposition process, the more predicted data will be unrecognized.

Table III compares the MSE of training and testing in each decomposition level to show the influence of data reduction, as the result of decomposition, to the accuracy. Prediction using the original LSTM RNN model without wavelet decomposition (Data S) generates the smallest error value compared to prediction using wavelet decomposition. This means that as the training data reduces as the result of the decomposition process, more data cannot be identified at the time of testing. The hybrid model at level 1 (A_1) generates an error value that is smaller than the hybrid model at level 2 (A_2) and 3 (A_3).

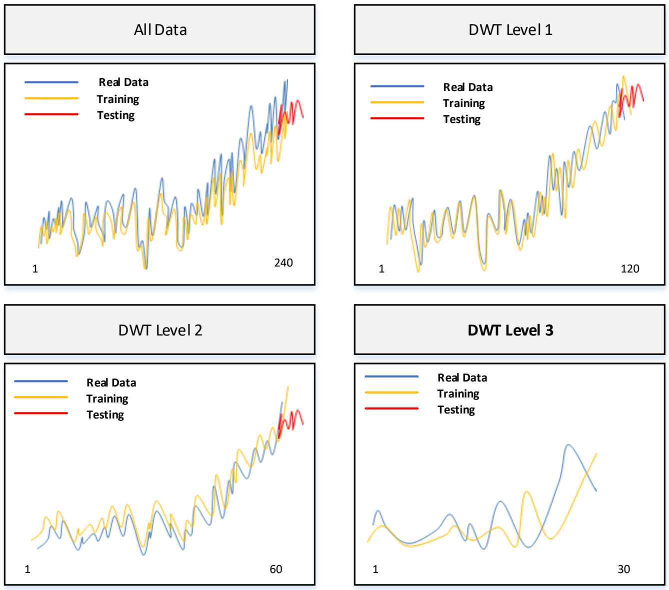


Fig. 8. Data training and data testing using different levels of decomposition

TABLE III. Comparison of Accuracy with Various Levels of Data Decomposition with MSE

Data	Training's MSE	Training Time	Testing's MSE
S	0.00464	00:59:11	0.02406
DWT level 1 (A_1)	0.11853	00:32:28	0.12870
DWT level 2 (A_2)	0.23689	00:14:16	0.02762
DWT level 3 (A_3)	0.44674	00:05:50	1.15110

From the experiments we have conducted, it can be inferred that the advantage of the use of the proposed model is mainly for shortening the time for data training rather than for increasing prediction accuracy. The lower accuracy compared to the original LSTM is due to the data reduction as the result of decomposition in wavelet transform. Hence, the proposed model can reduce the time of training process but not the error values.

IV. Conclusion and Future Work

In this paper, a hybrid model of wavelet transform and LSTM neural network is proposed to predict the number of tourist arrivals in Indonesia. This model incorporates wavelet and LSTM neural network to predict the number of tourist arrivals each month. The wavelet algorithm is used to decompose time series data into the data of low frequency and high frequency, which is proven to reduce the time for data training. The LSTM neural network is employed for training and testing the results of wavelet transform. The predicted outcome of the proposed hybrid model is compared to the original LSTM, Elman and Jordan RNN, as well as the hybrid of wavelet-Elman and the hybrid of wavelet-Jordan. The evaluation shows that the hybrid model of wavelet and LSTM method gives better training time than the original LSTM, Elman, and Jordan RNNs. Furthermore, this method is able to predict the number of tourist arrival more accurately than other hybrid methods.

One of the issues which is interesting for future work is to employ clustering method, such as k-means, to form the hybrid of LSTM RNN and clustering for time series prediction. The purpose of this is to compare the training time and the accuracy, to know which hybrid model can give better results.

REFERENCES

[1] The World Tourism Organization UNWTO, "International tourist arrivals up 4% in the first four months of 2015," *Press Release no. PR15048*, 2015. [Online]. Available: http://media.unwto.org/press-release/2015-07-08/international-tourist-arrivals-4-first-four-months-2015

[2] F. Murtagh, J.-L. Starck, and O. Renaud, "On neuro-wavelet modeling," *Decision Support System*, vol. 37, no. 4, pp. 475–484, September 2004.

[3] U. Lotric, "Wavelet based denoising integrated into multilayered perceptron," *Neurocomputing*, vol. 62, pp. 179–196, December 2004.

[4] L. Wang, K. K. Teo, and Z. Lin, "Predicting time series with wavelet packet neural networks," in *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, vol. 3, School of Electrical and Electronic Engineering nanyang Technological University. nanyang Avenue Singapore 639798: IEEE, 2001, pp. 1593–1597.

[5] X. Zhao, J. Lu, W. P. A. Putranto, and T. Yahagi, "Nonlinear time series prediction using wavelet networks with Kalman filter based algorithm," in *2005 IEEE International Conference on Industrial Technology*. IEEE, December 2005, pp. 1226–1230.

[6] G. Dominguez, M. Guevara, M. Mendoza, and J. Zamora, "A wavelet-based method for time series forecasting," in *2012 31st International Conference of the Chilean Computer Science Society*. IEEE, November 2012, pp. 91–94.

[7] L. F. Ortega, "A neuro-wavelet method for the forecasting of financial time series," in *International Conference on Soft Computing and Applications 2012*, vol. 1, WCECS. San Fransisco, USA: International Association of Engineers, October 2012, pp. 24–26.

[8] N. Terzija, "Robust digital image watermarking algorithms for copyright protection," Ph.D. dissertation, Universität Duisburg-Essen, January 2006.

[9] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 338–342.

[10] P. Fryzlewicz, S. V. Bellegem, and R. von Sachs, "Forecasting non-stationary time series by wavelet process modelling," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 4, pp. 737–764, December 2003.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[12] T. Thireou and M. Reczko, "Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 441–446, July 2007.

[13] Y. Tian and L. Pan, "Predicting short-term traffic flow by long short-term memory recurrent neural network," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, December 2015, pp. 153–158.

[14] K. Chen, Y. Zhou, and F. Dai, "A LSTM-based method for stock returns prediction: A case study of China stock market," in *2015 IEEE International Conference on Big Data (Big Data)*, October 2015, pp. 2823–2824.

[15] M. Shafie-khah, M. P. Moghaddam, and M. Sheikh-El-Eslami, "Price forecasting of day-ahead electricity markets using a hybrid forecast method," *Energy Conversion and Management*, vol. 52, no. 5, pp. 2165–2169, May 2011.

[16] Z. Zainuddin, N. Mahat, and Y. A. Hassan, "Improving the convergence of the backpropagation algorithm using local adaptive techniques," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 1, no. 1, pp. 184–187, December 2007.

[17] S. K. Panda and P. K. Jana, "Efficient task scheduling algorithms for heterogeneous multi-cloud environment," *The Journal of Supercomputing*, vol. 71, no. 4, pp. 1505–1533, January 2015.

[18] ——, "A multi-objective task scheduling algorithm for heterogeneous multi-cloud environment," in *2015 International Conference on Electronic Design, Computer Networks Automated Verification (EDCAV*. IEEE, January 2015, pp. 82–87.

[19] D. Sripathi, "Efficient implementations of discrete wavelet transforms using fpgas," Master's thesis, Florida State University, November 2003.

[20] T. W. Joo and S. B. Kim, "Time series forecasting based on wavelet filtering," *Expert Systems with Applications*, vol. 42, no. 8, pp. 3868–3874, May 2015.

[21] C. Stolojescu, I. Railean, S. Moga, P. Lenca, and A. Isar, "A wavelet based prediction method for time series," in *Proceedings of Stochastic Modeling Techniques and Data Analysis (SMTDA2010) International Conference, Chania, Greece*, 2010.

[22] K. Kawagoe and T. Ueda, "A similarity search method of time series data with combination of Fourier and wavelet transforms," in *Proceedings Ninth International Symposium on Temporal Representation and Reasoning*, 2002, pp. 86–92.

[23] J. Shi and L. Cheng, "Financial crisis dynamic prediction based on sliding window technology and Mahalanobis-Taguchi system," in *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, vol. 4, Sept 2011, pp. 65–68.

[24] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computing*, vol. 12, no. 10, pp. 2451–2471, October 2000.

[25] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2002.

[26] B. Rahmadi and Supriyadi, "Early model of traffic sign reminder based on neural network," *TELKOMNIKA*, vol. 10, no. 4, pp. 749–758, December 2012.

[27] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, pp. 1247–1250, Jun. 2014.

[28] M. Yousefi, D. Hooshyar, M. Yousefi, W. Khaksar, K. S. M. Sahari, and F. B. I. Alnaimi, "An artificial neural network hybrid with wavelet transform for short-term wind speed forecasting: A preliminary case study," in *2015 International Conference on Science in Information Technology (ICSITech)*. IEEE, October 2015, pp. 95–99.

Reverse Area Skyline in a Map

Annisa
Graduate School of Engineering,
Hiroshima University, Japan

Asif Zaman
Graduate School of Engineering,
Hiroshima University, Japan

Yasuhiko Morimoto
Graduate School of Engineering,
Hiroshima University, Japan

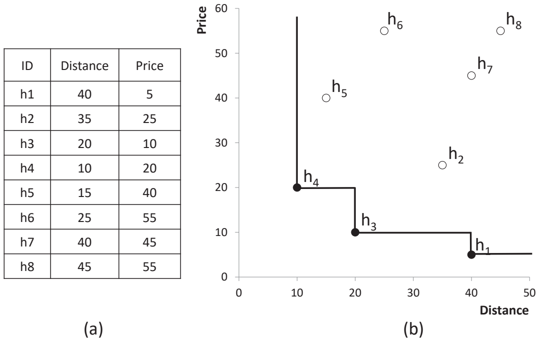
***Abstract*—Skyline query retrieves a set of data objects, each of which is not dominated by another object. On the other hand, given a query object q, "reverse" skyline query retrieves a set of points that are "dynamic" skyline of q. If q is a given preference of a user, "dynamic" skyline query retrieves a set of points that are not dominated by another point with respect to q. Intuitively,"reverse" skyline query of q retrieves a set of points that are as preferable as q. Area skyline query is a method for selecting good areas, each of which is near to desirable facilities such as stations, warehouses, promising customers' house, etc. and is far from undesirable facilities such as competitors' shops, noise sources, etc. In this paper, we applied reverse skyline concept to area skyline query and proposed Reverse Area Skyline algorithm. Analogically, given an area g, reverse area skyline query selects areas, each of which are as preferable as g. Assume that a real estate company wants to sell an area. Reverse area skyline query must be useful for such company to consider effective real estate developments so that the area attracts many buyers. Reverse area skyline query can also be used for selecting promising buyers of the area.**

***Keywords*—skyline query; reverse skyline query; area skyline query**

I. INTRODUCTION

Skyline query [1] is a widely applicable method for selecting small number of superior data objects. It retrieves a set of data objects, each of which is not dominated by another object. Given D as a d-dimensional database, an object p_i is said to dominate another object p_j if p_i is not worse in any of the d dimensions than p_j, and p_i is better than p_j in at least one of the d dimensions. Fig. 1 shows a typical example of skyline. Consider a typical online booking system. A user can select a hotel from the list in Fig. 1 (a) based on her/his preference on the price and distance of the hotel to the beach. Assume that smaller value is better for each attribute. In this situation, $\{h_1, h_3, h_4\}$ are skyline objects because they are not dominated by another object. Other objects $\{h_2, h_5, h_6, h_7, h_8\}$ are dominated by h_4. Fig. 1 (b) shows skyline hotels from the given hotel list.

Dynamic skyline query [2] and reverse skyline query [3] is a variant of skyline query. Given a query object q, "reverse" skyline query retrieves a set of points that are "dynamic" skyline of q. If q is a given preference of a user, "dynamic" skyline query retrieves a set of points that are not dominated by another point with respect to q. Intuitively,"reverse" skyline query of q retrieves a set of points that are as preferable as q. Assume that a businessman is running a hotel h_2. Reverse skyline query of h_2 retrieves a set of hotels that are as preferable as h_2. Therefore, he/she can expect customers who are interested in reverse skyline hotels of h_2, might also be interested in h_2. Skyline query retrieves candidate hotels from users' perspective, while reverse skyline query retrieves hotels from hotels' perspective.



ID	Distance	Price
h1	40	5
h2	35	25
h3	20	10
h4	10	20
h5	15	40
h6	25	55
h7	40	45
h8	45	55

Fig. 1: List of Hotel (a) and Conventional Skyline (b)

Skyline and reverse skyline query are two important methods for selecting smaller number of objects in various applications, one of which is in location selection problem. Choosing good location in a map is very important for many location-based applications. Usually, one would consider a location as a good location if it is near to desirable facilities that would be useful and/or pleasant to her/him such as stations, schools, supermarkets, etc. and is far from undesirable facilities that would unpleasant to her/him, such as competitors, noise sources, pollution sources, high-crime areas, etc.

Area skyline query is a method for selecting good areas, which are near to desirable facilities and far from undesirable facilities. In [4] and [5], the idea of skyline queries [1] is used to select area skyline in a map. We proposed Grid-based Area Skyline (GASky) algorithm in [5], which divides query area into grids as disjoint areas, and calculate minimum (min) and maximum (max) distance of each grid from closest desirable "+" facilities and closest undesirable "-" facilities. An area g dominates another area g' if g has smaller or equal max distance than min distance of g' for all facility types. Shaded grids in map in Fig. 2 shows an example of area skylines.

Area skyline query is important method to select non-dominated area from the users' perspective, who need some good locations based on his/her preference. Assume a real estate company has an area g (grid (2,18) in Fig. 2) to develop apartment, office, or market complex. The company needs to know who will be interested in the area. By using the idea of "reverse skyline", reverse skyline areas of g can be identified. Grey grids in Fig. 3 are dynamic area skyline of grid $(1, 14)$, while grey grids of Fig. 4 are reverse area skyline of g. Notice that g is a dynamic area skyline of grid $(1, 14)$, so that grid $(1, 14)$ is a reverse area skyline of g.

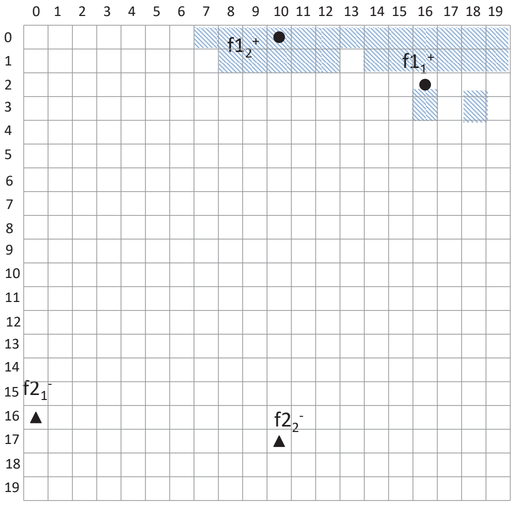


Fig. 2: Area Skyline Queries

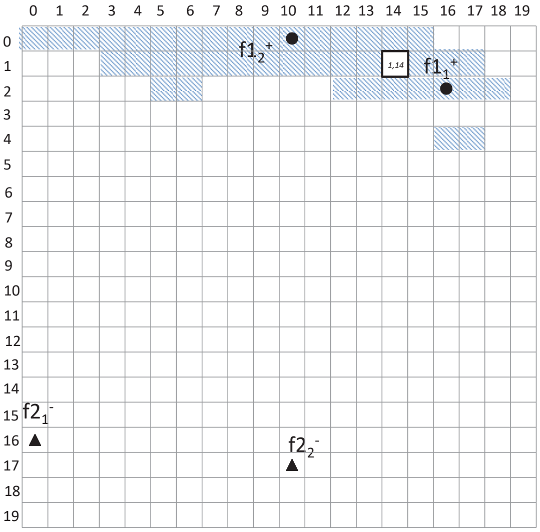


Fig. 3: Dynamic Area Skyline of grid (1,14)

On the analogy of the utilization of "reverse skyline", the "reverse" skyline areas has invaluable information. Let us consider a real estate company that have an area g (grid (2,18) in Fig. 4). Information about reverse area skyline, shaded area in Fig. 4, must be useful for such company to consider effective real estate developments so that the area attracts many buyers. Reverse area skyline query can also be used for selecting promising buyers of the area, since it may give the company clues to find who will be interested in the area. Moreover, it also may help to predict what type of business that would be suitable for the area considering the type of business that had already exist in the reverse area skylines.

In this paper, we present the reverse area skyline query and propose an effective and efficient method to answer reverse area skyline problem.

The contributions of this paper are summarized below:

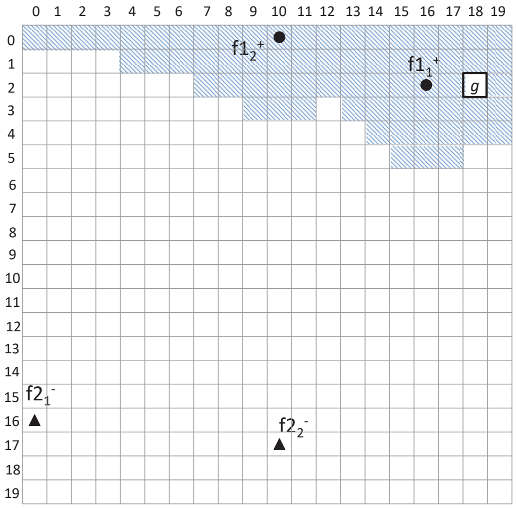


Fig. 4: Reverse Area Skyline Result

1) We have introduced a new skyline query, i.e., reverse area skyline query in the literature.
2) We have introduced some important concepts, dynamic area skyline and global area skyline, and propose Reverse Area Skyline (RASky) algorithm to answer the reverse area skyline problem.
3) We have conducted intensive experiments to prove the efficiency of proposed algorithm.

The rest of this paper is organized as follows. Section 2 reviews about skyline, dynamic skyline, reverse skyline, global skyline, spatial skyline, and area skyline issues. Section 3 formulates the problem definition and proposes the reverse area skyline algorithm. Section 4 presents the result of experiments, and finally Section 5 gives conclusions and future works.

II. Literature Review

A. Skyline, Dynamic Skyline, Reverse Skyline, and Global Skyline

Skyline query is a popular method for selecting small number of preferred answer from database. Since first introduced in [1], many algorithms have been proposed for answering skyline query problem, [1], [2], [6], [7]. Two important variants of skyline query are dynamic skyline query [2] and reverse skyline query [3]. Currently, the most efficient method in computing skyline and dynamic skyline is Branch and Bound Skyline (BBS), proposed in [2], which is a progressive algorithm using the R-tree, while The Branch and Bound Reverse Skyline (BBRS) algorithm [3] are the state-of-the-art algorithms for answering reverse skyline queries using the global skyline concept. In particular, BBRS is an improved customization of the original BBS algorithm [2]. GSRS is an improvement of BBRS to answer reverse skyline query [8].

In dynamic skyline query, a user specifies her/his preference as a query point in the data space and the query retrieves skyline objects that are not dominated by another with respect to the query object. In reverse skyline query, given a dataset P and a query point q in the space of P, an object p in P called as a "reverse skyline" of q if q is a dynamic skyline of p.

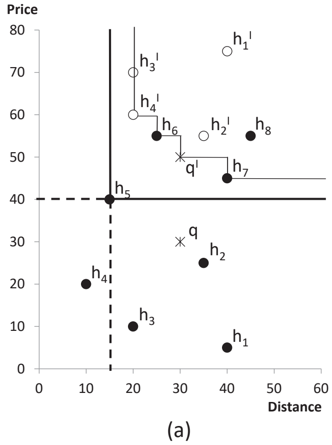


Fig. 5: Dynamic Skyline of h_5 (a) and Reverse Skyline (b)

Fig. 5 shows an example of a dynamic and reverse skyline queries. The example assumes that a user specifies $q = (30, 30)$ as a query point. To find reverse skyline of q, compute dynamic skyline of all objects, and find which objects that have q in its dynamic skyline. Fig. 5 (a) shows dynamic skyline query of h_5, $(x = distance, y = price) = (15, 40)$. To find dynamic skyline based on h_5, we first transform objects as follows. If x value of objects is less than 15, then transform the x value into $15 + (15 - x)$. Similarly, if y value of objects is less than 40, then transform y values into $40 + (40 - y)$. These transformations result in transformed data objects as in Fig. 5 (a). In the example, $h_2 = (35, 25)$ is transformed to $h'_2 = (35, 55)$, $h_4 = (10, 20)$ is transformed to $h'_4 = (20, 60)$, and so forth. We, then, compute skyline query for the transformed data objects as dynamic skyline query for $(15, 40)$, which retrieves $\{h_4, h_6, q, h_7\}$. Since q is in the dynamic skyline h_5, h_5 is a reverse skyline of q. Similarly, we calculate other reverse skyline objects of q as in Fig. 5 (b).

Skyline query in the Fig. 1 retrieves candidate hotels from users' perspective. On the other hand, reverse skyline query in Fig. 5 (b) retrieves hotels from hotels' perspective. Assume that a company is running a hotel whose detail is represented as a query point q. Intuitively, a user that is interested in h_5 hotel may also be interested in q since q is a dynamic skyline of h_5. The similar intuition holds on h_2 and h_6. Therefore, the company can expect users who are interested in h_2, h_5, and h_6 might also be interested in q.

Calculating dynamic skyline for each p in P to find reverse skyline of q needs very large computation. In order to reduce the search space, Dellis and Seeger introduced Branch and Bound Reverse Skyline (BBRS) algorithm using a concept called global skyline in [3].

Given a d-dimensional data set P and a query point q, p_1 said globally dominates p_2 with respect to q if: (1) (p_1-q)(p_2-q)>0 for all dimension, and (2) distance p_1 to q are smaller or equal than distance p_2 to q for all dimension, and smaller at least in one dimension. Rule (1) is to make sure that p_1 and p_2 are in the same quadrant w.r.t query point, while Rule (2) is dominance rule of global skyline. Evangelos and Seeger [3] have proved that reverse skyline point always a member of global skyline point.

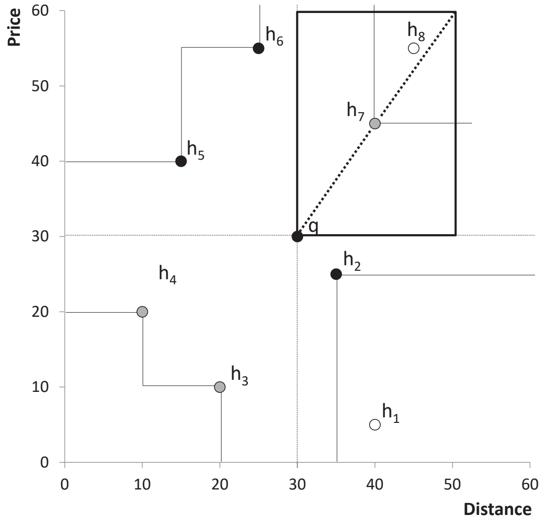


Fig. 6: Global Skyline and Query window

For selecting reverse skyline points from global skyline points, BBRS applies a window query for each global skyline point p. Window Query is an empty range query (boolean range query) which will return either true or false depending on whether there is any object inside the given range or not [9]. In [3], window query is the rectangle area with p as its center, and distance to a given query point q and its extension as its border coordinate. If there is another point inside this rectangle, then p is not a reverse skyline of q, otherwise p is a reverse skyline of q. Moreover, to reduce the number of window query checks, Gao et al. [8] introduced global-1-skyline concept in GSRS algorithm, which is a set of points that globally dominated by at most one global skyline point. They have proved that instead of checking all globally dominated points against all window queries, we simply just need to check whether any global skyline point or global 1-skyline point is inside the window query. Using the concept of global skyline point, BBRS and GSRS only consider points that potentially can be reverse skyline points and prune other points.

Fig. 6 shows an example of global skyline points of q ($\{h_2, h_3, h_4, h_5, h_6, h_7\}$), global 1-skyline points, h_1 and h_8 (white points), and the window query of h_7. Based on rule 1 in global skyline definition, each point is only compared with other points in the same quadrant. There are four quadrants in Fig. 6. h_5 and h_6 are in the same quadrant, h_7 and h_8, h_4 and h_3, and h_2 and h_1 are in the other quadrants. Since h_8 is inside the window query of h_7, then h_7 is not a reverse skyline of q. Applying window query to the rest of global skyline points, we can get the reverse skyline points of q are h_2, h_5, and h_6 (black points). h_7, h_3, and h_4 (grey points) are global skyline points, but not reverse skyline points since there is another point inside their window query.

Due to the importance of its use in various fields of application, research in the reverse skyline has gained many attention in the database research community such as in [3], [10]–[15]. All of the proposed reverse skyline variations above only consider about reverse skyline for zero dimensional data. Specifically, none of them considers about how to select reverse skyline from two dimensional objects, such as areas in a map. Therefore, all the previous algorithms can not directly be used to answer reverse area skyline problem.

B. Spatial Skyline Query

Spatial skyline query (SSQ) was first introduced in [16]. Given a set of data points P and a set of query points Q, an SSQ retrieves those points of P which are not dominated by any other point in P considering their derived spatial attributes, which is the point's distance to a query point. The difference between spatial skyline query with the regular skyline query is that the domination condition of P depends on the distance to query points Q.

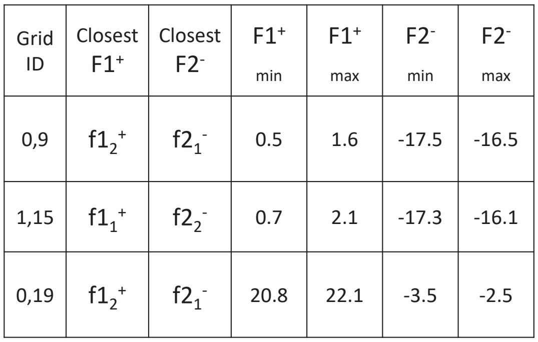
There are several researches works of spatial skyline problem, like in [17]–[23]. All of the above studies are based on the assumption that there are candidate points to choose skyline location and focused only on spatial data points, which is a zero dimensional data.

C. Area Skyline Query

Area skyline query was introduced in [4] and [5]. Given A as a domain area on map and g as rectangular query area in A. Let $F = \{F1, ...Fm\}$ be a set of facility types, which can be categorized into m types. Each type is classified into desirable (annotated by $+$ mark) or undesirable (annotated by $-$ mark). Each facility type has some number of facility objects m_i, for example, a desirable facility $F1^+$ has two objects $F1^+ = \{f1^+_1, f1^+_2\}$. Area skyline query using GASky algorithm consist of two steps. In the step one, GASky would divide A into sxt grids, where s is a number of rows and t is a number of columns. Let us consider a map in Fig. 2. Suppose a company would like to build a new housing complex in a region. To attract customers, the housing complex should be in an area that is near to train stations (point) and far from pollution source (triangle). There are two train stations $(f1^+_1, f1^+_2)$ and two pollution sources $(f2^-_1, f2^-_2)$ in this region. Note that "+" symbol is annotated to train stations, which are desirable facilities, and "-" symbol is annotated to pollution sources, which are undesirable facilities. In this situation, the company has to find two dimensional area on the map. In this example, the region is divided into 20×20 grids, say grid (0,0), ..., grid (19,19), each of which can be identified by row and column number. Then, GASky finds the closest train station and the closest pollution source from each grid and calculates min and max distance to the closest facilities and record the computation result into Minmax table like in Fig. 7. Min distance is the closest distance from grid to the closest facility, while max distance is the farthest distance from grid to the closest facility.

Since each grid is surrounded by four vertexes, to simplify in calculating min and max distance from each grid to closest facility type, GASky calculates distance from vertexes first. Using Voronoi diagram of each facility type, GASky finds the closest facility object for each type to each vertex, and then calculates its distance. Using vertex distance, GASky can calculate min and max distance for each grid easily [5]. For example, the closest $F1$ facility to four vertexes of a grid $g_{0,0}$ is $f1^+_2$ (as shown in Fig. 2). In common case, using these vertexes' distance information, GASky simply add the lowest value as min distance, and the highest as max distance. However, there are two special cases in calculating the min distance: first if the facility is inside the grid, and the other is if the facility is outside of the grid but one of the facility's coordinate is located between the coordinate of two vertexes. In the first case, 0 value simply added to the min distance of the grid, but in the second case we need to recalculate min distance from the facility to the edge connecting those two vertexes.

Fig. 7 is a partial Minmax table that records min and max distance to each facility type for grid $(0, 9)$, $(1, 15)$, and $(0, 19)$ in Fig. 2. In the table, notice that each distance value of undesirable facility was multiplied by -1 and swapped between its min and max value, so that we can say that the smaller value is better.



Grid ID	Closest $F1^+$	Closest $F2^-$	$F1^+$ min	$F1^+$ max	$F2^-$ min	$F2^-$ max
0,9	$f1_2^+$	$f2_1^-$	0.5	1.6	-17.5	-16.5
1,15	$f1_1^+$	$f2_2^-$	0.7	2.1	-17.3	-16.1
0,19	$f1_2^+$	$f2_1^-$	20.8	22.1	-3.5	-2.5

Fig. 7: Minmax Table

After completing Minmax table, in the second step GASky finds area skyline using area skyline dominance rule. A grid g would dominate another grid g' if for all distance to all facility type, max distance of g is smaller or equal than min distance of g'. Area skyline is a set of grids that are not dominated by another grid. Grid $(0, 9)$ and $(1, 15)$ dominates $(0, 19)$. Therefore, $(0, 19)$ is not an area skyline. GASky returns non-dominated areas (records) in the Minmax table (shaded area in Fig. 2) as area skyline.

The computational cost analysis of GASky step 1 shows that GASky takes $O(stm)$ in addition to the Voronoi diagrams' construction time, where s,t, and m are the number of rows, the number of columns, and the number of facility types, respectively. Experiment result in [5] shows that processing time of GASky increases when the number of facility type, the number of objects, and the number of grids increase. The ratio of skyline, which is the ratio between the number of skyline grids compared to all grids, would be high if the number of grids is small and the number of facility type and facility objects are big. If a user prefers small number of area skyline, she/he should increase the number of grids, so that the ratio of skyline areas will decrease. GASky can answer area selection based on user's perspective, but not from the company/business owner perspective. Nevertheless, since GASky operates on two dimensional object, we can use step one method of GASky to calculate Minmax table in reverse area skyline problem.

III. Reverse Area Skyline Query

In this section, we propose a reverse skyline query for grids in a map, which we call "Reverse Area Skyline Query".

A. Problem Definition

Let A be a rectangular target area in which there are spatial objects. Each spatial object can be categorized into one of m facility types. Let Fk be a set of type k ($k = 1, ...m$) objects, which are $Fk = \{fk_1, fk_2, ..., fk_{n_k}\}$ where n_k is the number of objects of the type k facility.

1) Grids and Vertexes: We divide A into $s \times t$ square grids where s is the number of rows, and t is the number of column. We can identify each grid using row number and column number. For example, $g_{i,j}$ is a grid that lies in the i-th row and the j-th column. Each square grid is surrounded by four vertexes, each of which can also be identified by row number and column number. For example, top-left, top-right, bottom-left, and bottom-right vertex of $g_{i,j}$ can be identified as $v_{i,j}$, $v_{i,j+1}$, $v_{i+1,j}$, and $v_{i+1,j+1}$, respectively. In Fig. 8, we defined the query grid g, and divided an area into 12x12 grids. $g_{4,6}$ is surrounded by four vertexes $v_{4,6}$, $v_{4,7}$, $v_{5,6}$ and $v_{5,7}$, respectively. For each vertexes, we find the nearest object of each facility type using Voronoi diagram. After that, we calculate min and max distance from each grid using the same calculation in GASky step 1 as discussed in Section II-C, and record the distances in the Minmax table. From now, we call one record in Minmax table as one object.

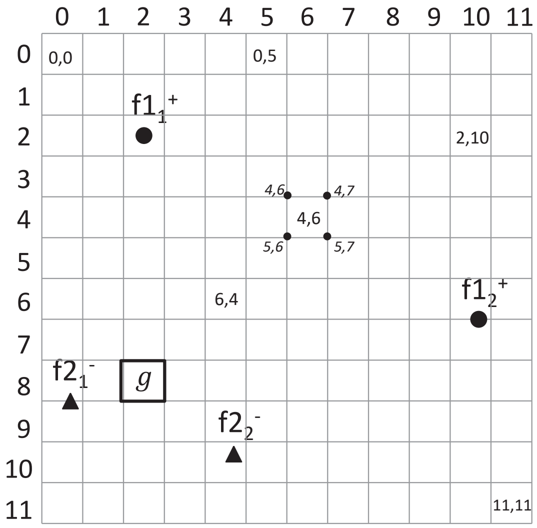


Fig. 8: Target area divided into 12 x 12 grids

2) Dynamic Area Skyline: Let $min(d_k(g))$ and $max(d_k(g))$ be the the min and max distance to facility f_k of grid g, and $min(d_k(q))$ and $max(d_k(q))$ be the the min and max distance to facility f_k of query grid q. In order to calculate dynamic area skyline, we need to transform the distances similar to conventional dynamic skyline. Let $min(d_k(g))^T$ and $max(d_k(g))^T$ are the transformed min and max distance, respectively. There are six cases to transform $min(d_k(g))$ and $max(d_k(g))$ w.r.t query grid q into $min(d_k(g))^T$ and $max(d_k(g))^T$. Fig. 9 illustrates the transformation of six cases in dynamic area skyline.

In all cases, we assume $min(d_k(q))$ and $max(d_k(q))$ are 5 and 8, respectively. In case 1, assume $min(d_k(g))$ and $max(d_k(g))$ are 9 and 11. Since 9 and 11 are larger than 8, then $min(d_k(q))^T$ and $max(d_k(q))^T$ become 1 (9-8) and 3 (11-8). In case 2, assume $min(d_k(g))$ and $max(d_k(g))$ are 6 and 11. Since 6 is between 5 and 8, and 11 is larger than 8, then $min(d_k(q))^T$ and $max(d_k(q))^T$ become 0 and 3 (11-8). In case 3, assume $min(d_k(g))$ and $max(d_k(g))$ are 6 and 7. Since 6 and 7 are between 5 and 8, then $min(d_k(q))^T$ and $max(d_k(q))^T$ become 0. In case 4, assume $min(d_k(g))$ and $max(d_k(g))$ are 1 and 11. Since 1 is smaller than 5 and 11 is larger and 8, then $min(d_k(q))^T$ and $max(d_k(q))^T$ become 0 and 3, just like in case 2. In case 5, assume $min(d_k(g))$ and $max(d_k(g))$ are 1 and 6. Since 1 is smaller than 5 and 6 is between 5 and 8, then $min(d_k(q))^T$ and $max(d_k(q))^T$ become 0 and 4 (5-1). In case 6, assume $min(d_k(g))$ and $max(d_k(g))$ are 1 and 4. Since 1 and 4 are smaller than 5, then $min(d_k(q))^T$ and $max(d_k(q))^T$ become 1 (5-4) and 4 (5-1).

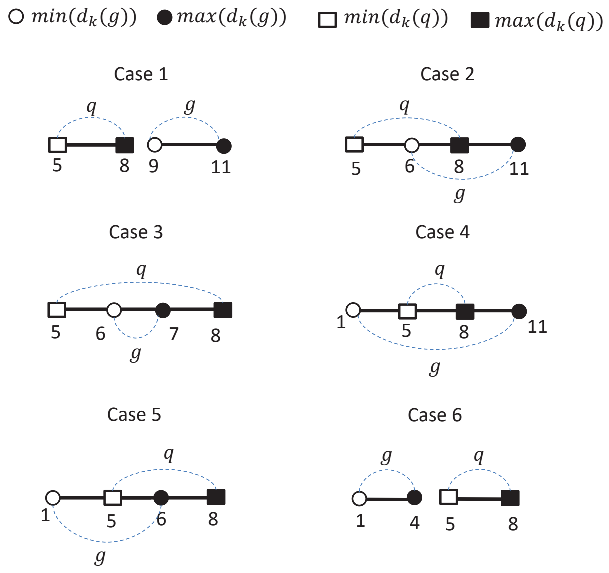


Fig. 9: Transformation cases

We then formally defined Case 1 to 6 as:

if $min(d_k(g)) \geq max(d_k(q))$, then

$$\begin{aligned} min(d_k(g))^T &= min(d_k(g)) - max(d_k(q)) \\ max(d_k(g))^T &= max(d_k(g)) - max(d_k(q)) \end{aligned} \tag{1}$$

if $max(d_k(g)) > max(d_k(q))$ and $max(d_k(q)) > min(d_k(g)) \geq min(d_k(q))$, then

$$\begin{aligned} min(d_k(g))^T &= 0 \\ max(d_k(g))^T &= max(d_k(g)) - max(d_k(q)) \end{aligned} \tag{2}$$

if $min(d_k(g)) \geq min(d_k(q))$ and $max(d_k(g)) \leq max(d_k(q))$, then

$$\begin{aligned} min(d_k(g))^T &= 0 \\ max(d_k(g))^T &= 0 \end{aligned} \tag{3}$$

if $min(d_k(g)) < min(d_k(q))$ and $max(d_k(g)) > max(d_k(q))$, then

$$\begin{aligned} min(d_k(g))^T &= 0 \\ max(d_k(g))^T &= max(d_k(g)) - max(d_k(q)) \end{aligned} \tag{4}$$

if $min(d_k(g)) < min(d_k(q))$ and $min(d_k(q)) < max(d_k(g)) \leq max(d_k(q))$, then

$$\begin{aligned} min(d_k(g))^T &= 0 \\ max(d_k(g))^T &= min(d_k(q)) - min(d_k(g)) \end{aligned} \tag{5}$$

if $max(d_k(g)) \leq min(d_k(q))$, then

$$\begin{aligned} min(d_k(g))^T &= min(d_k(q)) - max(d_k(g)) \\ max(d_k(g))^T &= min(d_k(q)) - min(d_k(g)) \end{aligned} \tag{6}$$

Definition 1. Dynamic Area Skyline Query

For two objects, g and g', we said g dynamically dominates g' w.r.t q, if and only if $max(d_k(g))^T \leq min(d_k(g'))^T$ for all k ($1 \leq k \leq m$). Dynamic area skyline query of q retrieves the set of all area objects that are not dynamically dominated by any other objects w.r.t q.

Based on dynamic area skyline definition, we can formally define the reverse area skyline of query area q.

Definition 2. Reverse Area Skyline Query

Let G be a set of d-dimensional objects. Reverse area skyline query w.r.t query area q retrieves all area objects $g \in G$ where q is in the dynamic area skyline of g. In other words, we said g is reverse area skyline of q if $\nexists g' \in G$ such that $max(d_k(g'))^T \leq min(d_k(q))^T$ for all k ($1 \leq k \leq m$) w.r.t g.

Using Definition 2, we can compute reverse area skyline query by performing dynamic area skyline query for each grid objects, and retrieve set of grid objects which have query area q in their dynamic area skyline result. But as discussed in Section II-A, to compute reverse skyline by computing dynamic skyline for each object is time-consuming. In this paper, we define global area skyline concept to compute reverse area skyline. We extend global skyline concept in [3] so that it can be applied in area skyline.

3) Disjoint, Overlap, Within/Contain: Using information of min and max distances, one object's min and max distance might disjoint, overlap, within/contain with another object's min and max distance. Let us consider the example of Fig. 9 again. We define disjoint objects using case 1 and 6, overlap objects using case 2 and 5, and case 3 and 4 for within/contain objects.

Definition 3. Disjoint Objects

Object g is disjoint with g', if $max(d_k(g)) \leq min(d_k(g'))$ or $min(d_k(g)) \geq max(d_k(g'))$, for all $k \in m$.

Definition 4. Overlap Objects

Object g overlap with g', if $max(d_k(g)) > max(d_k(g'))$ and $max(d_k(g')) > min(d_k(g)) \geq min(d_k(g'))$, or if $min(d_k(g)) < min(d_k(g'))$ and $min(d_k(g')) < max(d_k(g)) \leq max(d_k(g'))$, for at least one $k \in m$.

Definition 5. Within/Contain Objects

Object g is within g', if $min(d_k(g)) \geq min(d_k(g'))$ and $max(d_k(g)) \leq max(d_k(g'))$, for all $k \in m$. Object g contains g' if $min(d_k(g)) < min(d_k(g'))$ and $max(d_k(g)) > max(d_k(g'))$, for all $k \in m$.

These disjoint, overlap, and within/contain conditions are two dimensional objects' characteristics that are not exist in zero dimensional objects. Based on these characteristics, we define Lemma 1 and 2 which are very important to efficiently compute reverse area skyline using global area skyline concept.

Lemma 1. *Let q be the query area. If g overlaps with or within/contain q, then g must be a reverse area skyline of q.*

Proof. Assume g is not a reverse area skyline of q. Then, there should be at least one object that dynamically dominates q w.r.t g. If we apply dynamic area skyline of g, since g overlaps or within/contains q, based on case 2, 3, 4, 5 in Section III-A2, $min(d_k(g))^T$ is always be 0, which makes it not possible to be dominated by other objects. It means that q is a dynamic area skyline of g, and consequently, g is reverse area skyline of q. So the assumption is not true and the proof is complete.◇

Fig. 10 shows an illustration of Lemma 1. Fig. 10 (a) shows original min-max distance of q and g, while Fig. 10 (b) shows min-max distance of q^T after we apply dynamic area skyline of g.

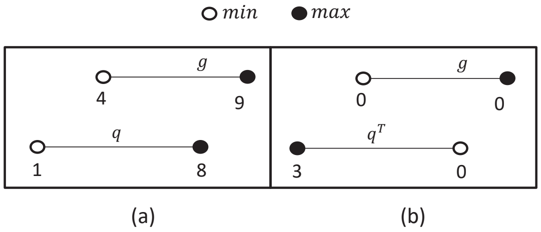


Fig. 10: Lemma 1 situation

Lemma 1 provides an easy selection method for RASky algorithms to directly put overlap/within/contain objects into reverse area skyline result.

Before defining global area skyline, let us consider the definition of global skyline for zero dimensional data in [3]. Given a d-dimensional data set P and a query point q, p_1 is said globally dominates p_2 with respect to q if: (1) (p_1-q)(p_2-q)>0 for all dimension, and (2) distance p_1 to q are smaller or equal then distance p_2 to q for all dimension, and smaller at least in one dimension. Let us consider point h_2 and h_1 in Fig. 6. We said h_2 is globally dominates h_1, because (h_1-q)(h_2-q)>0 and distance h_2 to q are smaller than distance h_1 to q for all dimension. In two dimensional case, the above situations become more complicated for overlap/within/contain conditions. Lemma 2 shows that overlap/within/contain objects can not globally dominate any other objects.

Lemma 2. *Overlap/within/contain objects can not globally dominate any other objects*

Proof. Let assume that g and g' is in the same quadrant w.r.t q. g is an overlap/within/contain object w.r.t q, and g' has $min(d_k(g'))^T \geq max(d_k(g))^T$, for all $k \in m$, so that g globally dominated g', and g' is not a reverse area skyline of q. On the contrary, if we apply dynamic area skyline of g', we can see that g can not dominate q, since g and q are overlap/within/contain objects. This mean that g' is reverse skyline of q and should not be discarded by g.◇

Fig. 11 shows an illustration of Lemma 2. Fig. 11 (a) shows the original min-max distance of q, g, and g'. Fig. 11 (b) shows if g (overlap/within/contain object) globally dominate g' w.r.t q, since $max(d_k(g))^T$ is smaller than $min(d_k(g'))^T$, then g' is not a reverse skyline of q. But this will lead to wrong result. If we apply dynamic area skyline on g' as in Fig. 11 (c), q is in dynamic area skyline of g', which consequently makes g' as reverse area skyline of q. In this situation g should not be allowed to globally dominate g' at the first place, since g' is reverse area skyline of q. Using Lemma 1 and 2, we can reduce the comparison step in calculating global area skyline because all the overlap/within/contain objects do not participate in the comparison process.

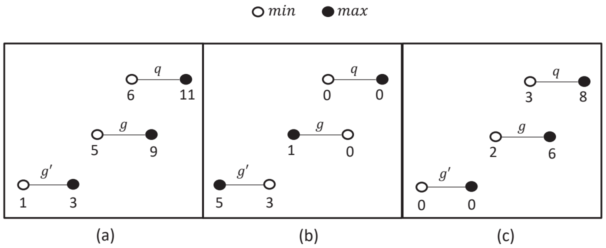


Fig. 11: Lemma 2 situation

4) Global and Global-1 Area Skyline: Based on Lemma 1 and 2, only disjoint objects will participate in global area skyline computation. Let us consider disjoint situations in Fig. 9 case 1 and 6. In Fig. 9 case 1, $min(d_k(g))$ and $max(d_k(g))$ are larger than $max(d_k(q))$, while in Fig. 9 case 6 $min(d_k(g))$ and $max(d_k(g))$ are smaller than $min(d_k(q))$. To differentiate between these two disjoint conditions, we defined $\mathit{diff}(g_k)$ as:

$$\mathit{diff}(g_k) = \begin{cases} min(d_k(g)) - max(d_k(q)) \\ \qquad\qquad \text{if } max(d_k(q)) \le min(d_k(g)). \\ max(d_k(g)) - min(d_k(q)) \\ \qquad\qquad \text{if } max(d_k(g)) \le min(d_k(q)). \end{cases}$$

Notice that the value of $\mathit{diff}(g_k)$ could be “positive” (case 1) or “negative” (case 6). Two objects g and g' are in the same quadrant w.r.t q if $(\mathit{diff}(g_k))(\mathit{diff}(g'_k)) > 0$ for all $k \in m$. In Fig. 12, since $max(d_k(g)) \le min(d_k(q))$,(15 $\le$ 20), then $\mathit{diff}(g_k) < 0$ while $\mathit{diff}(g'_k) > 0$, since $min(d_k(g')) \ge max(d_k(q))$, (30 $\ge$ 25). Using Lemma 1, Lemma 2, and *diff* definition, we define global and global-1 area skyline.

Definition 6. Global and Global-1 Area Skyline For two objects, g and g', we said g globally dominates g' w.r.t q, if and only if: (1)g and g' are disjoint objects w.r.t q, (2)$(\mathit{diff}(g_k))(\mathit{diff}(g'_k)) > 0$ and (3) $max(d_k(g))^T \le min(d_k(g'))^T$, for all $k \in m$. Any objects g'' becomes global-1 area skyline if there is only one other object that globally dominates it.

5) Window Query: Window Query of grid $w(g)$ w.r.t q, has minimum and maximum value for each k dimension, $min(w_k(g))$ and $max(w_k(g))$ where $k \in m$. It is defined as follows:

$$min(w_k(g)) = \begin{cases} max(d_k(q)) \qquad \text{if } min(d_k(g)) \ge max(d_k(q)). \\ min(d_k(g)) + \mathit{diff}(g_k) \\ \qquad\qquad \text{if } max(d_k(g)) \le min(d_k(q)). \end{cases}$$

$$max(w_k(g)) = \begin{cases} max(d_k(g)) + \mathit{diff}(g_k) \\ \qquad\qquad \text{if } min(d_k(g)) \ge max(d_k(q)). \\ min(d_k(q)) \qquad \text{if } max(d_k(g)) \le min(d_k(q)). \end{cases}$$

Fig. 12 shows an illustration of window query’s minimum and maximum value in one dimension. Assume min and max distance for g, q, and g' are (10,15), (20,25), and (30,35). For $w(g)$, since $max(d_k(g)) \le min(d_k(q))$, then $\mathit{diff}(g_k)$ is -5 (15-20), so that $min(w_k(g))$ and $max(w_k(g))$ are 5 (10 + (-5)) and 20 (same value as $min(d_k(q))$). For $w(g')$, since $min(d_k(g)) \ge max(d_k(q))$, then $\mathit{diff}(g'_k)$ is 5 (30-25), so that $min(w_k(g'))$ and $max(w_k(g'))$ are 25 (same value as $max(d_k(q))$) and 40 (35+5).

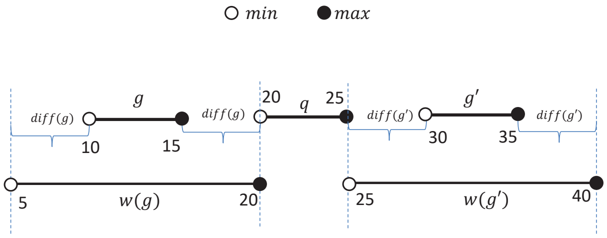


Fig. 12: Diff and Window query

Lemma 3. *Let g be a global area skyline of q, and g' be a global or global 1-area skyline of q with the same quadrant with g. If the window query of g contains g' w.r.t q, then g is not a reverse area skyline of q.*

Proof. If the window query of g contains g', then if we apply dynamic area skyline of g using formula in Section III-A2, we know that $max(d_k(g'))^T$ is always smaller than $min(d_k(q))^T$. It means that g' will dynamically dominate q w.r.t g, therefore g can not be a reverse area skyline of q.$\diamond$

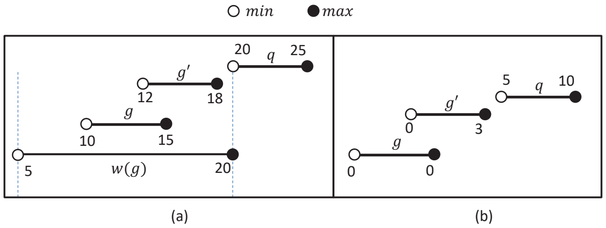
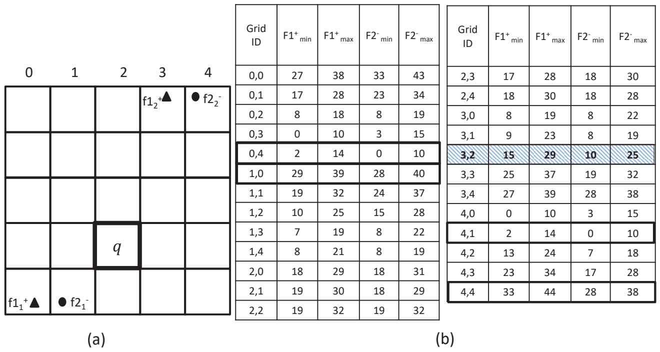


Fig. 13: Lemma 3 situation

Fig. 13 illustrates Lemma 3 situation. Fig. 13 (a) shows that $w(g)$ is contain g', while Fig. 13 (b) shows that g' will dynamically dominate q w.r.t g, so that g is not a reverse area skyline of q. Using Lemma 3, for each global area skyline we simply just check whether at least one of other global or global 1-area skyline is within its window query or not.

B. Reverse Area Skyline (RASky) Algorithm

Reverse area skyline algorithm (RASky) consist of two steps. At step 1, we divide A into grids. For each grid, we find the nearest facility type, calculate its min and max distance, and complete the distance information in Minmax table using the same method in GASky step 1 [5] as explained in Section II-C. In step 2, using information in Minmax table from the first step, we calculate reverse area skyline using global area skyline. In this section, we will focus on the reverse area skyline step 2.



Grid ID	$F1^+_{min}$	$F1^+_{max}$	$F2^-_{min}$	$F2^-_{max}$
0,0	27	38	33	43
0,1	17	28	23	34
0,2	8	18	8	19
0,3	0	10	3	15
0,4	2	14	0	10
1,0	29	39	28	40
1,1	19	32	24	37
1,2	10	25	15	28
1,3	7	19	8	22
1,4	8	21	8	19
2,0	18	29	18	31
2,1	19	30	18	29
2,2	19	32	19	32

Grid ID	$F1^+_{min}$	$F1^+_{max}$	$F2^-_{min}$	$F2^-_{max}$
2,3	17	28	18	30
2,4	18	30	18	28
3,0	8	19	8	22
3,1	9	23	8	19
3,2	**15**	**29**	**10**	**25**
3,3	25	37	19	32
3,4	27	39	28	38
4,0	0	10	3	15
4,1	2	14	0	10
4,2	13	24	7	18
4,3	23	34	17	28
4,4	33	44	28	38

Fig. 14: Sample map (a) and Minmax Table (b)

In this section we use sample map in Fig. 14 (a) and set $g_{3,2}$ as query area q, then divide sample map into 5x5 grids. In this map we have two types of facilities, $F1^+$ and $F2^-$, each of them have two objects $F1^+ = (f1_1^+, f1_2^+)$, $F2^- = (f2_1^-, f2_2^-)$. After completing RASky step 1 in the sample map, we obtain Minmax table like in Fig. 14 (b).

We index the grid by their min and max distance in Minmax table using R-tree structure. Each leaf in the R-tree is in the format $(id, qd, RECT)$, where id is the number of grid in Minmax table, qd is quadrant, and $RECT$ is a bundle of all min and max distance in a grid for all dimension. For example, for 2 facility type, or 2-dimensional, $RECT$ has $(f1min, f2min)$ as bottom-left coordinate and $(f1max, f2max)$ as top-right coordinate. Our query object $RECT_{3,2}$ has bottom-left coordinate (15,10) and top-right coordinate (29,25). Using $RECT$ object, we build R-tree of Minmax table.

1) Building R-tree: RASky read each object in Minmax table. Using Lemma 1 and 2, if the object is an overlap/within/contain object, then it will automatically be a reverse area skyline object, and will be excluded from R-tree and further computation. Only disjoint objects will be inserted into R-tree. Let us consider Minmax table in Fig. 14 (b). Since $g_{0,4}$, $g_{1,0}$, $g_{4,1}$, and $g_{4,4}$ are disjoint objects (rows with bold border in Fig. 14 (b)), they are inserted into R-tree, while others directly become reverse area skyline of q. Fig. 15 shows R-tree after inserting disjoint objects.

2) Finding Global and Global-1 area skyline: RASky insert all root entries into heap H and sort them by their distance from q. Besides H, we also use two additional heap H_g and H_{g1} to maintain global area skyline and global-1 area skyline. Since $N1$ is closest to q, its entry is expanded, and $N1$ is removed from H. Now H contents become $RECT_{0,4}$, $RECT_{4,1}$, and $N2$. As top of H, $RECT_{0,4}$ then become the first global area skyline and inserted into H_g. Notice that $RECT_{4,1}$ is in the same quadrant with $RECT_{0,4}$ and it is not globally dominated by $RECT_{0,4}$, so it also inserted into H_g. Next $N2$ is expanded and $RECT_{1,0}$ and $RECT_{4,4}$ is inserted into H. Since $RECT_{1,0}$ is in different quadrant with $RECT_{0,4}$ and $RECT_{4,1}$, $RECT_{1,0}$ also become global area skyline and inserted into H_g. $RECT_{4,4}$ is in the same quadrant with $RECT_{1,0}$, but since it is not globally dominated by $RECT_{1,0}$, then it also inserted into H_g. Since there is no global-1 area skyline in this sample dataset, then H_{g1} is remain empty.

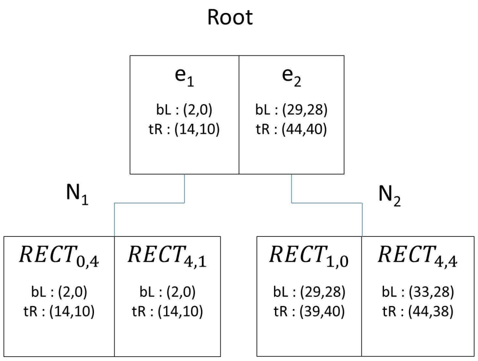
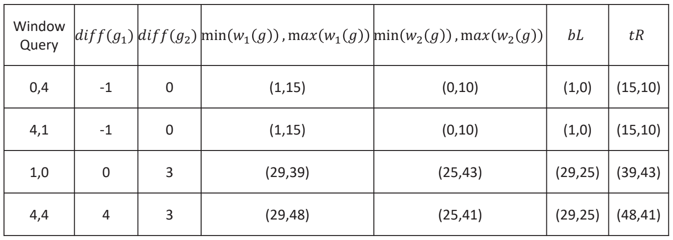


Fig. 15: R-tree of disjoint objects

3) Applying Window Query: After getting all global area skyline, we build window query for each entry in H_g. Using window query formula in Section III-A5, Fig. 16 shows bottom-left and top-right coordinate of each window query for query area $RECT_{3,2}$ whose bottom-left and top-right is (15,10) and (29,25), respectively.



Window Query	$diff(g_1)$	$diff(g_2)$	$\min(w_1(g)), max(w_1(g))$	$\min(w_2(g)), max(w_2(g))$	bL	tR
0,4	-1	0	(1,15)	(0,10)	(1,0)	(15,10)
4,1	-1	0	(1,15)	(0,10)	(1,0)	(15,10)
1,0	0	3	(29,39)	(25,43)	(29,25)	(39,43)
4,4	4	3	(29,48)	(25,41)	(29,25)	(48,41)

Fig. 16: Window query in sample dataset

Let us consider $w(g_{0,4})$, $diff(g_{0,4_1})$ is -1 (14-15) and $diff(g_{0,4_2})$ is 0 (10-10). Using these values, we can compute min and max of $w(g_{0,4})$ in dimension 1 as (2+(-1),15) and in dimension 2 as (0+0,10), so that bL and tR coordinates are (1,0) and (15,10). Since $RECT_{4,1}$ has the same bL and tR coordinate with $RECT_{0,4}$, then min and max of $w(g_{4,1})$ are the same with $w(g_{0,4})$. This mean that $RECT_{4,1}$ always contains $w(g_{0,4})$ and vice versa, so based on Lemma 3, both of them are not reverse area skyline of $g_{3,2}$. Now for $w(g_{1,0})$, $diff(g_{1,0_1})$ is 0 (29-29) and $diff(g_{1,0_2})$ is 3 (28-25). Min and max of $w(g_{1,0})$ in dimension 1 as (29,39+0) and in dimension 2 as (25,40+3), so that bL and tR coordinates are (29,25) and (39,43). Finally, for $w(g_{4,4})$, $diff(g_{4,4_1})$ is 4 (33-29) and $diff(g_{4,4_2})$ is 3 (28-25). Min and max of $w(g_{4,4})$ in dimension 1 as (29,44+4) and in dimension 2 as (25,38+3), so that bottom-left and top-right coordinates are (29,25) and (48,41). $RECT_{4,4}$ overlaps with $w(g_{1,0})$, while window query of $w(g_{4,4})$ contains $RECT_{1,0}$. Based on Lemma 3, $RECT_{1,0}$ is reverse area skyline of q while $RECT_{4,4}$ is not. From the above computation, we can find that $g_{0,4}$, $g_{4,1}$, and $g_{4,4}$ is not reverse area skyline of $g_{3,2}$ while the others are.

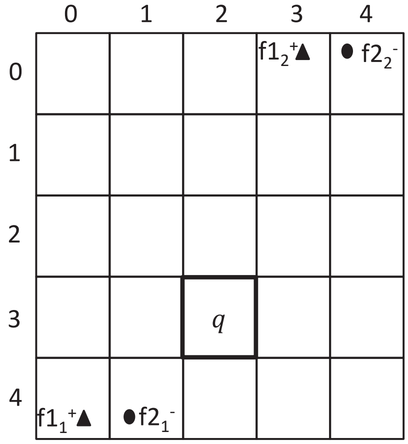


Fig. 17: Reverse area skyline for sample map

Shaded area in Fig. 17 shows reverse area skyline of $g_{3,2}$ in sample map Fig. 14 (a), which is 88% of all grids. Next in the experiment section, we discover that smaller size of query area q will reduce reverse area skyline result.

IV. Experimental Evaluation

We experimentally evaluated RASky algorithm in a PC with Intel Core i5 3.2GHz processor and 4GB of RAM. We conducted three experiments using three synthetic datasets. In each experiment, we repeated five times and reported the average. We examined the effect of parameters such as number of objects, number of types, and number of grids, to the step 1 and step 2 of RASky algorithm. We recorded the processing time for step 1 and step 2 of RASky and the ratio of reverse area skyline resulted from each experiment. Ratio of skyline is the number of reverse area skyline compared with the number of grids in the experiment, Table I lists the synthetic datasets and parameters in these experiments.

TABLE I: Experimental Dataset

Dataset	Objects	Types	Grids
DB1	1k,2k,4k,8k,16k	2	160k
DB2	1k	2,4,8,16	40k
DB3	1k	2	10k,40k,160k,640k

A. Effect of Number of Objects

In these experiments, we examined the performance of RASky on the different number of objects, when the number of facility types and the number of grids are fix, using DB1. Fig. 18 shows the processing time of this experiment. We can see that the increase of the number of objects will increase the total processing time of RASky. In step 1, increasing number of objects will increase the processing time to build Voronoi diagram. However, since the number of Voronoi diagram is fix according to the number of type, increasing number of objects does not have effect in the size of Minmax table. Hence in RASky step 2, increasing the number of objects has less effect, and the processing time tend to decrease when the number of objects increases. The reason is because increasing number of objects, while the number of grids is fix, increases the number of non-disjoint objects. It means less objects will participate in global area skyline computation, since only disjoint objects participates in the computation. Therefore the processing time will be decreased. The ratio of reverse area skyline are increasing as reported in Fig. 19. Increasing the number of objects will cause smaller value on min and max distances, but since the number of grids is fix when the number of objects increase, the ratio of reverse area skyline still will increase.

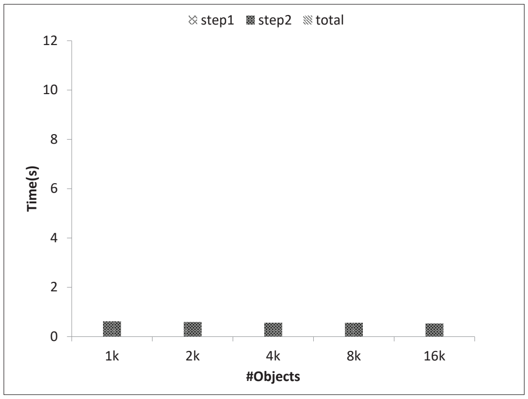


Fig. 18: Processing time of $DB1$

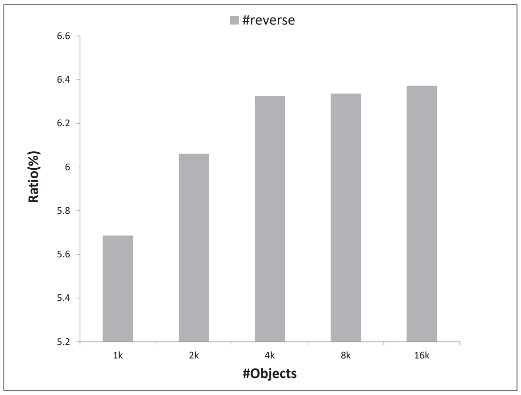


Fig. 19: Reverse area skyline's ratio of $DB1$

B. Effect of Number of Types

In these experiments, we used a synthetic data DB2 that have fix number of objects and number of grids. From the results in Fig. 20, we can observe that the processing time increases with the increase of the number of types. The increasing number of types will require more Voronoi diagrams, which in turn increase the processing time. The result illustrates that increasing the number of types significantly increase the processing time of step 1. Similar with increasing number of objects, increasing number of types with fix number of grids will decrease the number of disjoint objects. Although the number of disjoint objects decreases, the processing time still increases because increasing facility types also means larger size of Minmax tables. Since the dimension is increasing as the number of facility types increase, the ratio of skyline is also increasing as shown in Fig. 21.

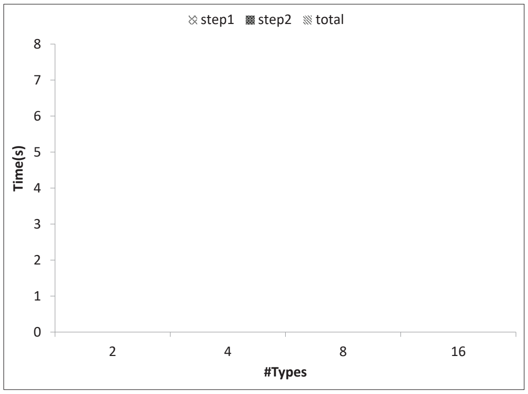


Fig. 20: Processing time of $DB2$

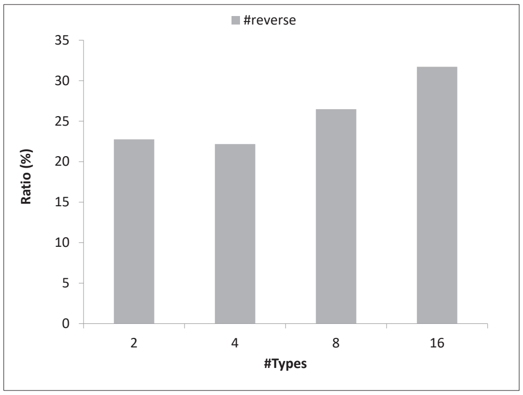


Fig. 21: Reverse area skyline's ratio of $DB2$

C. Effect of Number of Grids

In these experiments, we evaluated the effect of number of grids while the number of objects and number of types are fix, using DB3. Fig. 22 shows that the number of grids affects the processing time of step 1 and step 2. In step 1, increasing the number of grids means more comparison on Voronoi diagrams and more calculation time to obtain min and max distance, and in the same time enlarges the number of record in Minmax table which also cause increasing time needed for step 2 computation. Increasing number of grids while number of objects and number of types are fix also causing the number of disjoint objects to increase. In step 2, increasing number of disjoint objects will increase processing time since more objects will participate in global area skyline computation. In Fig. 23, the effect of number of grids affect ratio of skyline differently compared to the effect of the number of objects and types. Increasing the number of grids will decrease the ratio of skyline. The important reason of that is because more grids has the same meaning of having smaller size of each grid, which significantly decrease the ratio of skyline, since smaller area is likely to be dominated by another area.

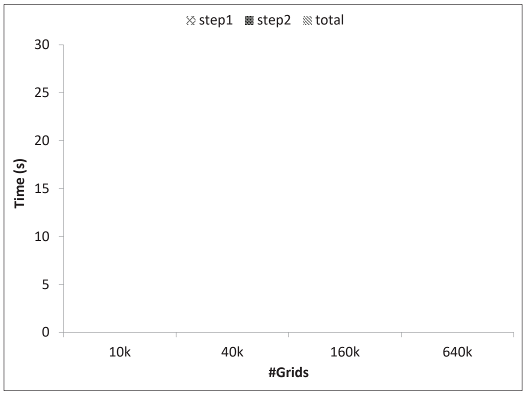


Fig. 22: Processing time of $DB3$

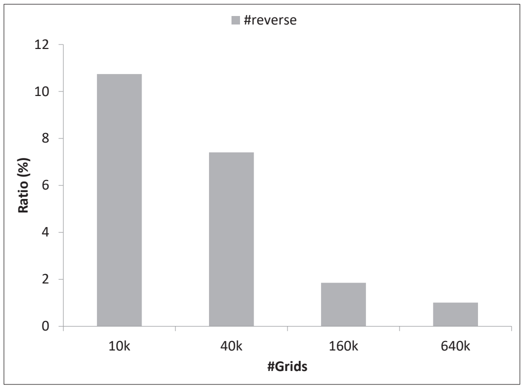


Fig. 23: Reverse area skyline's ratio of $DB3$

From all of experimental results, we can indicate that the total processing time of RASky increases when the number of objects, number of facility types, and the number of grids increases. In addition, the ratio of skyline increases when the number of objects and types increases, and decreases when the number of grids increases.

V. Conclusions and Future Works

In this paper, we define dynamic area skyline, global area skyline, and propose reverse area skyline algorithm (RASky) to solve the reverse area skyline query. This query is very important for location selection in business' or landowners' perspective. RASky has two steps, step 1 to compute Minmax table and step 2 to calculate reverse area skyline. Smaller query area will obtain smaller number of reverse area skyline and vice versa. Reverse area skyline gives invaluable information for landowner to pursue targeted customer or to decide what type of business that would attract more customer. Comprehensive experiments are conducted to show the effectiveness and efficiency of the proposed algorithms. In the future, we will consider another skyline problem in two dimensional objects, such as selecting k-dominant areas. We are also interested in the application of this method to road network, which also taken into account nonspatial properties such as population density, price, traffic condition, and so on.

Acknowledgment

The authors would like to thank KAKENHI (16K00155) Japan for supporting this research, Indonesian Government DG-RSTHE scholarship for supporting Annisa, and Japanese Government MEXT Scholarship for supporting Asif Zaman.

References

[1] S. Borzsonyi, D. Kossmann, and K. Stocker, "The skyline operator," in *Proceedings of the 17th International Conference on Data Engineering (ICDE), April 2–6, Heidelberg, Germany*, 2001, pp. 421–430.

[2] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "An optimal and progressive algorithm for skyline queries," in *Proceedings of the ACM SIGMOD June 9–12, California, USA*, 2003, pp. 467–478.

[3] D. Evangelos and B. Seeger, "Efficient computation of reverse skyline queries," in *Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment*, 2007, pp. 291–302.

[4] Annisa, M. A. Siddique, A. Zaman, and Y. Morimoto, "A method for selecting desirable unfixed shape areas from integrated geographic information system," in *Proceedings of IIAI*, 2015, pp. 195–200.

[5] Annisa, A. Zaman, and Y. Morimoto, "Area skyline query for selecting good locations in a map," *Information Processing Society of Japan : Database Transaction*, vol. 9, no. 3, pp. 0–0, 2016.

[6] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, "Skyline with Presorting," in *Proceedings of the 19th International Conference on Data Engineering (ICDE), March 5–8, Bangalore, India*, 2003, pp. 717–719.

[7] K.-L. Tan, P.-K. Eng, and B. C. Ooi, "Efficient progressive skyline computation," in *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), September 11–14, Rome, Italy*, 2001, pp. 301–310.

[8] Y. Gao, Q. Liu, B. Zheng, and G. Chen, "On efficient reverse skyline query processing," *Expert Systems with Applications*, vol. 40, no. 7, pp. 3237–3249, 2014.

[9] H. F. Singh, Amit and A. aman Tosun, "High dimensional reverse nearest neighbor queries," in *Proceedings of the twelfth International Conference on Information and Knowledge Management*, 2003, pp. 91–98.

[10] W. Xiaobing, Y. Tao, R. C.-W. Wong, L. Ding, and J. X. Yu., "Finding the influence set through skylines," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, ACM*, 2009, pp. 1030–1041.

[11] L. Zhu, C. Li, and H. Chen, "Efficient computation of reverse skyline on data stream," in *Computational Sciences and Optimization, International Joint Conference on, vol. 1, IEEE*, 2009, pp. 735–739.

[12] G. Wang, J. Xin, L. Chen, and Y. Liu, "Energy-efficient reverse skyline query processing over wireless sensor networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 7, pp. 1259–1275, 2012.

[13] Q. Liu, Y. Gao, G. Chen, Q. Li, and T. Jiang, "On efficient reverse k-skyband query processing," in *International Conference on Database Systems for Advanced Applications*, 2012, pp. 544–559.

[14] Y. Park, J.-K. Min, and K. Shim, "Parallel computation of skyline and reverse skyline queries using mapreduce," in *Proceedings of the VLDB Endowment 6, no. 14*, 2013, pp. 2002–2013.

[15] M. S. Islam, R. Zhou, and C. Liu, "On answering why-not questions in reverse skyline queries," in *IEEE 29th International Conference on Data Engineering*, 2013, pp. 973–984.

[16] M. Sharifzadeh and C. Shahabi, "The spatial skyline queries," in *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), September 12–15, Seoul, Korea*, 2006, pp. 751–762.

[17] K. Kodama, Y. Iijima, X. Guo, and Y. Ishikawa, "Skyline queries based on user locations and preferences for making location-based recommendations," in *Proceedings of the International Workshop on Location Based Social Networks (LBSN) November 03, Washington, USA*, 2009, pp. 9–16.

[18] M. Arefin, J. Xu, Z. Chen, and Y. Morimoto, "Skyline query for selecting spatial objects by utilizing surrounding objects," *Journal of Software*, vol. 8, no. 7, pp. 1742–1749, 2013.

[19] X. Guo, Y. Ishikawa, and Y. Gao, "Direction-based spatial skylines," in *Proceedings of the 9th ACM International Workshop on Data Engineering for Wireless and Mobile Access (MobiDE), June 6, Indiana, USA*, 2010, pp. 73–80.

[20] Q. Lin, Y. Zhang, W. Zhang, and X. Lin, "Efficient general spatial skyline computation," *World Wide Web*, vol. 16, no. 3, pp. 247–270, 2013.

[21] G.-W. You, M.-W. Lee, H. Im, and S.-W. Hwang, "The farthest spatial skyline queries," *Information Systems*, vol. 38, no. 3, pp. 286–301, 2013.

[22] Y.-W. Lin, E.-T. Wang, C.-F. Chiang, and A. L. P. Chen, "Finding targets with the nearest favor neighbor and farthest disfavor neighbor by a skyline query," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC), March 24–28, Gyeongju, Korea*, 2014, pp. 821–826.

[23] M. Arefin, G. Ma, and Y. Morimoto, "A spatial skyline query for a group of users," *Journal of Software*, vol. 9, no. 11, pp. 2938–2947, 2014.

OpenCL-Accelerated Object Classification in Video Streams using Spatial Pooler of Hierarchical Temporal Memory

Maciej Wielgosz
AGH University of Science and Technology
Kraków, Poland

Marcin Pietroń
Academic Computer Centre CYFRONET,
of the University of Science and Technology in Cracow
Kraków, Poland

***Abstract*—The paper presents a method to classify objects in video streams using a brain-inspired Hierarchical Temporal Memory (HTM) algorithm. Object classification is a challenging task where humans still significantly outperform machine learning algorithms due to their unique capabilities. A system which achieves very promising performance in terms of recognition accuracy have been implemented. Unfortunately, conducting more advanced experiments is very computationally demanding; some of the trials run on a standard CPU may take as long as several days for 960x540 video streams frames. Therefore, authors decided to accelerate selected parts of the system using OpenCL. In particular, authors seek to determine to what extent porting selected and computationally demanding parts of a core may speed up calculations.**

The classification accuracy of the system was examined through a series of experiments and the performance was given in terms of F1 score as a function of the number of columns, synapses, *min_overlap* and *winners_set_size*. The system achieves the highest F1 score of 0.95 and 0.91 for *min_overlap*=4 and 256 synapses, respectively. Authors have also conduced a series of experiments with different hardware setups and measured CPU/GPU acceleration. The best kernel speed-up of 632x and 207x was reached for 256 synapses and 1024 columns. However, overall acceleration including transfer time was significantly lower and amounted to 6.5x and 3.2x for the same setup.

***Keywords*—Hierarchical Temporal Memory; OpenCL; GPU; Video processing**

I. Introduction

Despite the huge technological growth witnessed nowadays, there are still no autonomous machines available which would be capable of operating in the real world. Such machines would take over most of our tedious everyday duties and clear the way for a breakthrough in Artificial Intelligence. However, such robots need to be able to process inputs in real time, learn, generalize and react to events. This requires building an appropriate processing system which has human–like capabilities.

A mammalian brain is an example of such a system which evolved over millions of years. Despite its apparent complexity there is only one algorithm [1] within the brain which governs the body functions. This allows for scalability of the solutions based on the algorithm since more complex systems may be built on a top of the simpler ones just by duplication of the basic structure.

The human brain as a whole has not been completely explored yet, making its artificial implementation and verification a very hard task. However, there are initiatives [2] which have taken up the challenge of simulating and modeling a brain as we know it today. Rather than model the brain, the authors of this paper have adopted a slightly different approach of gradually introducing selected components of Hierarchical Temporal Memory (HTM) to the video processing system with the intention of enhancing its performance. By doing so, authors aim to develop a complete system [3] working on the principles of the human brain as they were presented in [1], [4] with necessary modifications making the algorithm suitable for hardware implementation. Running HTM on CPU is very slow and the algorithm due to its strongly parallel structure is a good candidate for General–Purpose Graphics Processing Unit (GPGPU) and Field–Programmable Gate Array (FPGA) acceleration. Consequently, this paper presents an architecture of GPU implementation of Spatial Pooler (SP). The computationally demanding overlap and inhibition sections of SP were implemented on GPU.

The rest of the paper is organized as follows. Sections I-A and I-B provide the background and related work of Hierarchical Temporal Memory and object classification in video streams, respectively. The data flow in the custom–designed system used for the experiments is presented in Section II with system architecture described in Section III. Section IV provides the results of the experiments. Finally, the conclusions of conducted research are presented in Section V.

A. Hierarchical Temporal Memory

Hierarchical Temporal Memory (HTM) replicates the structural and algorithmic properties of the neocortex. It can be regarded as a memory system which is not programmed, but trained through exposing it to data flow. The process of training is similar to the way humans learn which, in its essence, is about finding latent causes in the acquired content. At the beginning, the HTM has no knowledge of the data stream causes it examines, but through a learning process it explores the causes and captures them in its structure. The training is considered complete when all the latent causes of data are captured and stable. The detailed presentation of HTM is provided in [4]–[6].

HTM constitutes a hierarchy of nodes, where each node performs the same algorithm. The most basic elements (raw

```
for all col $\in$ sp.columns do
    col.overlap $\leftarrow 0$
    for all syn $\in$ col.connected_synapses() do
        col.overlap $\leftarrow$ col.overlap + syn.active()
    end for
    if col.overlap $<$ min_overlap then
        col.overlap $\leftarrow 0$
    else
        col.overlap $\leftarrow$ col.overlap * col.boost
    end if
end for
```

Fig. 1. Overlap algorithm

```
for all col $\in$ sp.columns do
    max_column $\leftarrow$ max(n_max_overlap(col, n), 1)
    if col.overlap $>$ max_column then
        col.active $\leftarrow 1$
    else
        col.active $\leftarrow 0$
    end if
end for
```

Fig. 2. Inhibition algorithm

and unprocessed data) enter at the bottom of the hierarchy. Each node learns the spatio–temporal pattern of its input and associates it with a given concept. Consequently, each node, no matter where it is in the hierarchy, discovers the causes of its input. In an HTM, beliefs exist at all levels in the hierarchy and are internal states of each node. They represent probabilities that a cause is active. Each node in an HTM has a fixed number of concepts and a fixed number of output variables. The training process of an HTM starts with a fixed number of possible causes, and in a training process, assigns a meaning to them.

Consequently, the nodes do not increase the number of concepts they cover; instead, over the course of the training, the meaning of the outputs gradually changes. This happens at all levels in the hierarchy simultaneously. Thus the top level of the hierarchy remains with little or no meaning till nodes at the bottom are trained to recognize the basic patterns.

HTM is composed of two main parts, namely Spatial and Temporal Pooler (TP). This paper focuses on Spatial Pooler (SP), aka Pattern Memory, which is employed in the processing flow of the system. It contains columns with synapses connected to the input data [4]. The main role of SP in HTM is finding spatial patterns in the input data. It may be decomposed into three stages:

- Overlap calculation (Fig. 1),
- Inhibition (Fig. 2),
- Learning.

The first two stages are very computationally demanding but can be parallelized. Therefore the authors decided to implement them on GPU in OpenCL. The learning stage, the detailed description of which is provided in the Numenta whitepaper [4], is implemented on CPU.

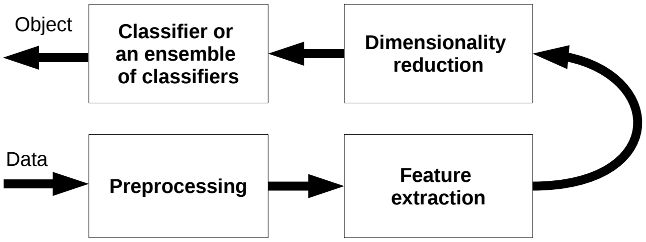


Fig. 3. Architecture of a video processing system

The overlap section (Fig. 1) computes $col.overlap$ for every column in SP structure i.e. a number of active and connected synapses. If the number is larger than $col.min\_overlap$, then it is boosted and passed on to the inhibition section (Fig. 2).

The inhibition stage (Fig. 2) implements a winner–takes–all procedure where for each column a decision is made as to whether it belongs to a range of n ($winners\_set\_size$) columns of the highest values. The $n\_max\_overlap()$ function performs the comparison.

B. Object classification in video streams

Most state–of–the–art information extraction systems consist of the following sections: preprocessing, feature extraction, dimensionality reduction and classifier or ensemble of classifiers (Fig. 3). Their construction requires expert knowledge as well as familiarity with the data that will be processed [7], [8].

Usually, systems for object classification in video streams are also designed according to this scheme. Consequently, the proper choice of the operations which constitute all the mentioned stages of the system is important and determines the classification result [9]–[11]. One of the most challenging stages is feature extraction, which substantially affects the overall performance of the system.

There are also systems which take advantage of the spatial–temporal [4] profile of the data [12]–[15]. They are closer to the concept of the solution presented in this paper, which may be considered a hybrid approach since it features components of both schemes.

II. Processing Flow

The data is fed into the system in a frame–by–frame manner. In the first step, the original frame is turned into a binary image (see III-A2). This conversion constitutes the encoding which allows the generation of input data for the SP processing stage.

Thereafter, the encoded data is fed into the SP. The processing done by the SP effectively maps input to Sparse Distributed Representation (SDR), which then may be passed on to the TP. The TP is not used in this particular application, but the system in general has such a capability. Instead, the TP is substituted with histograms to serve a similar purpose.

Histograms of consecutive frames are built from SP output on a per–video basis. The histograms are used as the input

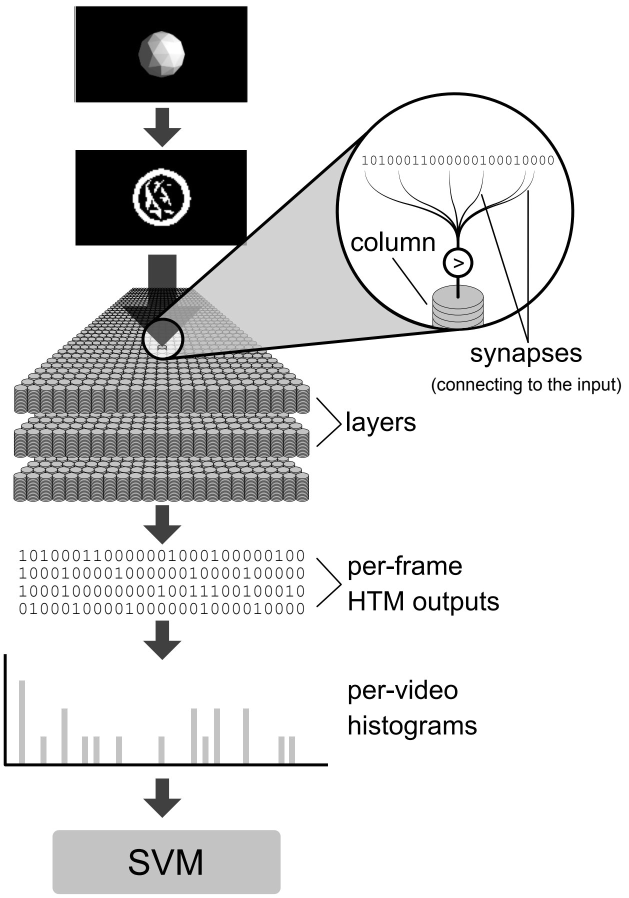


Fig. 4. Block diagram of the proposed approach

data for the SVM classifier which comes next. Classifier maps the results from SDR to the result space (output categories).

The complete processing flow of the system is presented in Fig. 4.

III. System Description

The system is highly configurable, with numerous parameters responsible for the core HTM's structure, the encoder behavior, statistics rendering, etc. The configuration is stored in a file written in JSON format, which allows it to maintain its readability while providing a clear structure. In addition to the core module, a set of supporting modules has been developed. Most of them are used for feeding video data to the core module, and receiving and analyzing the results.

The HTM itself is a 'core' module, in addition to the ones necessary for the system to function (responsible for data reading and encoding, as well as results interpretation) and ones created for debugging and statistics gathering purposes. The overall system architecture is depicted in Fig. 5. The most relevant modules are described in detail below.

A. Outer Structure

The outermost level of system is CLI (Command Line Interface). Depending on the provided command line options, it invokes a particular setup – either 'Single HTM' or 'Multiple HTMs'. In the 'Single HTM' setup data from all categories is fed into a single HTM instance. 'Multiple HTMs' refers to creating HTM instances on a per–category basis, resulting in an ensemble of one–vs–all detectors.

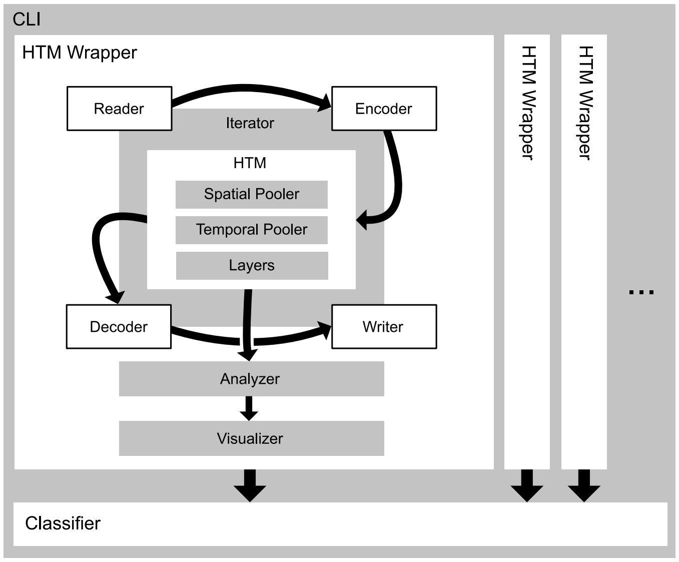


Fig. 5. Architecture of the implemented system

In both modes the same wrappers encapsulating the actual processing units can be used. A wrapper is created for a particular HTM use – it is responsible for creating relevant data readers, encoders, decoders and output writers, and for passing them to the iterator – a part of the core that manages HTM cycles.

After data is processed by the wrapper, the result reaches CLI, which is responsible for further analysis and data presentation – combining wrappers outputs, gathering statistics, training the classifier used to provide the final results, rendering data visualizations etc. The HTM results are post-processed using a LinearSVM classifier.

1) HTM Wrapper: As mentioned above, a wrapper is created for a specific use – the one designed to work with videos will differ from the one tailored for texts. Assembling a wrapper from predefined or newly created modules is the main task of the experiment setup.

The wrapper used in the present system setup creates a reader able to get data from video files and an encoder that converts raw frame data to the required format. The HTM output is neither modified (a pass-through decoder module) nor stored for future reference (a pass-through writer module).

Preparing the processing units to work is not the wrapper's only responsibility – it also controls the number of executed iterations. The minimum (and default) number of cycles equals a single pass of the learning set, however setups specifying maximum number and/or metrics measuring whether HTM still needs learning are also possible.

The wrapper module also coordinates statistics gathering and visualization on a per-instance basis.

2) Adaptive Video Encoder: During the encoding process an original video frame is converted to a binary image. Depending on the configuration, the original image can be first reduced in size to trim down the amount of data. After reduction, the color image is converted to a grayscale one, which is later binarized using adaptive thresholding.

Adaptive thresholding uses a potentially different threshold value for each small image region. It gives better results than using a single threshold value for images with varying illumination. In this encoder 'ADAPTIVE_THRESH_-GAUSSIAN_C' algorithm from OpenCV library [16] is used – a threshold value is the weighted sum of neighbourhood values where weights are a gaussian window.

B. HTM Core

All implemented readers, encoders, decoders and writers provide pre-defined interfaces. Such a solution allows to separate data acquisition and output storage from the actual processing. The loop consisting of a data retrieval, processing and outputting is executed by the iterator object of the core module.

1) HTM: An HTM object itself consists of a configurable number of layers, a Spatial Pooler and a Temporal Pooler object. Upon each iteration, each layer state is updated by SP and (depending on the configuration) TP, based on the data it receives. In the case of the lowest layer the input is obtained from the encoder, and for the higher ones – from the previous level. Setting the layer number to zero effectively turns off the HTM, causing the whole module's output to be equal to that of the encoder. This feature was used when comparing performance of 'SVM' only with the 'SP + SVM' ensemble.

Layers consist of columns, which are composed of connectors (containing synapses used in the spatial pooling process) and cells (used in temporal pooling). Cells themselves are built from segments, with each segment containing synapses connecting it to the other cells. This hierarchical structure closely mirrors the one described in the algorithm section.

Every object encapsulates its functionality, making introduction of changes and enhancements trivial, while at the same time providing a clear reference point for modifications. The object-oriented structure also enhances the visibility of a very important HTM feature – its potential for massive parallelization. One example of that can be a spatial pooling process. The initial system setup used a sequential version of SP. After some tests, a decision to replace it with a concurrent implementation running on a GPU (and an FPGA in the future) was made. The replacement spatial pooler, taking advantage of OpenCL capabilities, was written and plugged into the system without changes to the rest of the architecture.

2) Hardware architecture: The overlap calculation is a computationally intensive operation, executed multiple times for every input. Fig. 6 presents the hardware architecture of the overlap unit which was implemented in OpenCL. The main idea behind the presented architecture is based on a concept of locating each column in a separate GPU block (work group). This enables parallel calculation of each column's overlap which is only limited by global–to–local memory data transfer. Once the data is available in the local memory of each work group, a reduction operation is initiated. Intermediate results are stored in the local memory, and in the last stage the results from each block are sent over to the global memory of the GPU. It is worth noting (Fig. 6) that the boost operation [4] is also computed by each kernel within the work group.

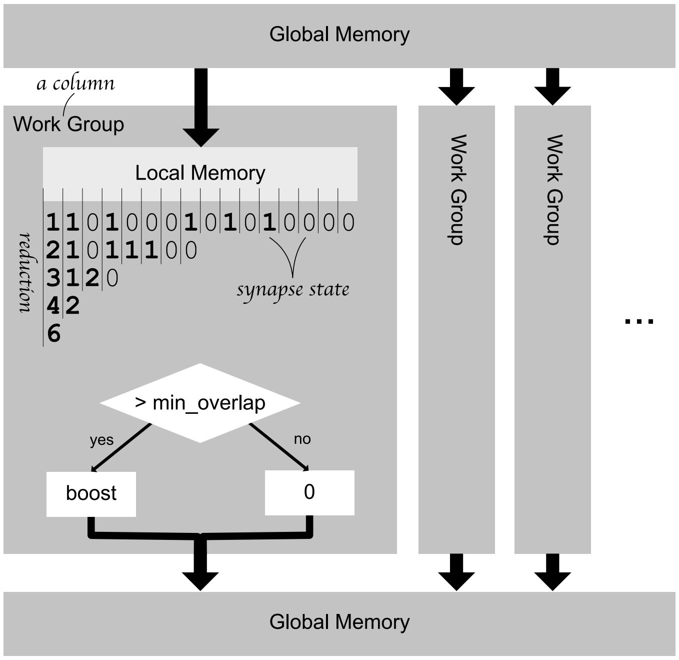


Fig. 6. Overlap implemented in OpenCL

The inhibition section presented in Fig. 7 may be considered as an extension of the overlap kernel. It builds up on top of the overlap kernel. The results of the overlap operation are sent back to the global memory of GPU to be fetched again to GPU blocks during the inhibition calculation procedure. The amount of the data required by every work group depends on the inhibition radius. When the overlap data are collected in each work group, a reduction, summation operation and $winners\_set\_size$ comparison is performed. The last operation directly affects the column state by changing it to active or inactive. Extending the overlap module with the logic related to the inhibition calculation improved the performance gain of system as presented in Fig. 18.

IV. Experiments and the Discussion

This section presents both quality assessment and acceleration results of the video classification system. It is worth noting that the output of CPU and GPU implementation is not exactly the same due to random initialization of the HTM parameters (e.g. synapses $init\_perm$ values) and learning/testing sets randomization.

All the tests presented in this chapter were performed on Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz with Radeon R9 390 STRIX GPU platform and 32 GB DD3 1600 MHz memory.

A. Experiments setup

A series of experiments (details of which are provided in Tab. I and Tab. II) was conducted. The experiments allow

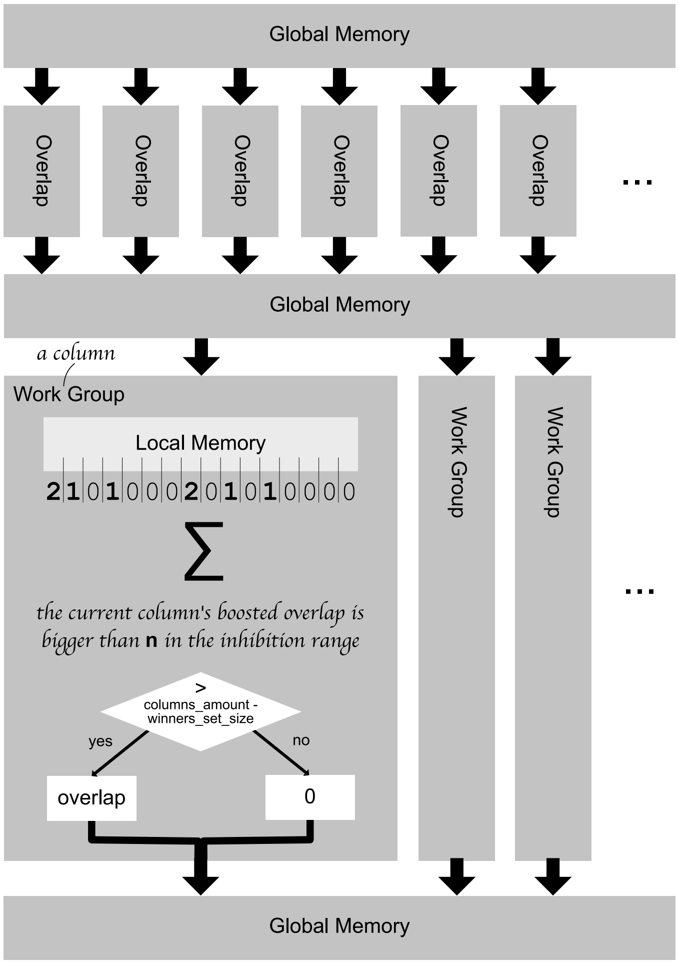


Fig. 7. Overlap + Inhibition implemented in OpenCL

TABLE I. EXPERIMENTS DETAILS

Size of a single video frame		240x134
No. of frames in a single video		32
Object classes		cone, cube, cylinder, monkey, sphere, torus
No. of classes		6
Total no. of videos	all	6000
	training	4800
	testing	1200
Videos per class	all	1000
	training	800
	testing	200
Videos per trial	all	100
	training	80
	testing	20

TABLE II. BASIC CONFIGURATION PARAMETERS

No. of columns	2048
No. of synapses per column	128
Perm value increment	0.1
Perm value decrement	0.1
Min overlap	8
Winners set size	40
Initial perm value	0.21
Initial inhibition radius	80

to compare the performance of the system featuring Spatial Pooler in the processing flow with the one lacking it, and to measure execution times of both implementations on CPU and GPU.

The experiments were conducted using a 'Single HTM' setup (see III-A). For each trial, the system was trained in the learning mode with 80% of available data (80 videos of each class randomly selected from a pool of 800) and then was tested with the remaining 20% of the data in the testing mode (20 videos per class selected out of 200).

During the course of an experiment the value of a single configuration parameter was changed, while the rest remained as in Tab. II. Each generated configuration was then used to run tests both on GPU and CPU using OpenCL inhibition kernel. Additionally, the same experiments with columns and synapses were conducted also for the overlap kernel (Fig. 18).

B. Dataset

The challenging part involved generation of sample videos for testing. The videos had to meet a series of requirements such as object location, camera location and object–camera distance. Consequently, a dedicated application was used to generate the videos (i.e. Blender [17]). Original rendered videos had a size of 960x540 pixels and showed a single, centered, stationary object with camera moving around it (Fig. 8).

For the experiments, the dataset (available online [18]) based on the rendered videos was created, with the frame resized to 240x134 pixels. The initial testing showed that reducing the frame size has a very small impact on SVM results (used as a baseline for comparison), while significantly shortening the HTM calculation time.

C. Quality assessment

The F1 score is used as a quality evaluation of the experiments' results presented in this paper. The precision and recall for corresponding clusters are calculated as follows:

$$Recall(i,j) = \frac{n_{ij}}{n_i}, \tag{1}$$

$$Precision(i,j) = \frac{n_{ij}}{n_j}, \tag{2}$$

where n_{ij} is the number of items of class i that are classified as members of cluster j, while n_j and n_i are the numbers of items in cluster j and class i, respectively. The cluster's F1 score is given by the following formula:

$$F(i,j) = 2 \cdot \frac{Recall(i,j)Precision(i,j)}{Precision(i,j) + Recall(i,j)}. \tag{3}$$

The overall quality of the classification can be obtained by taking the weighted average F1 scores for each class. It is given by the equation:

$$F1 = \sum_i \frac{n_i}{n} maxF(i,j), \tag{4}$$

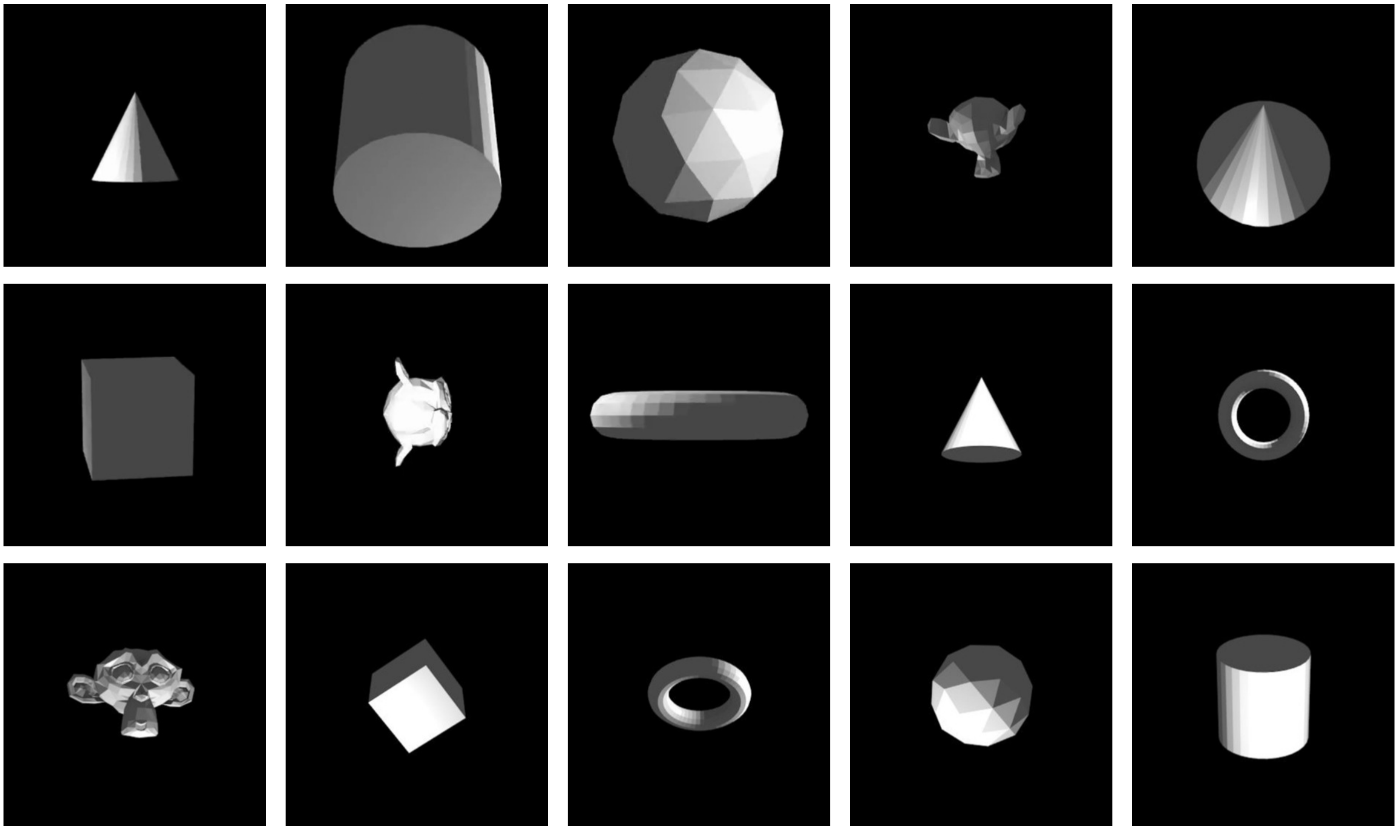


Fig. 8. Sample frames of different shapes rendered in Blender

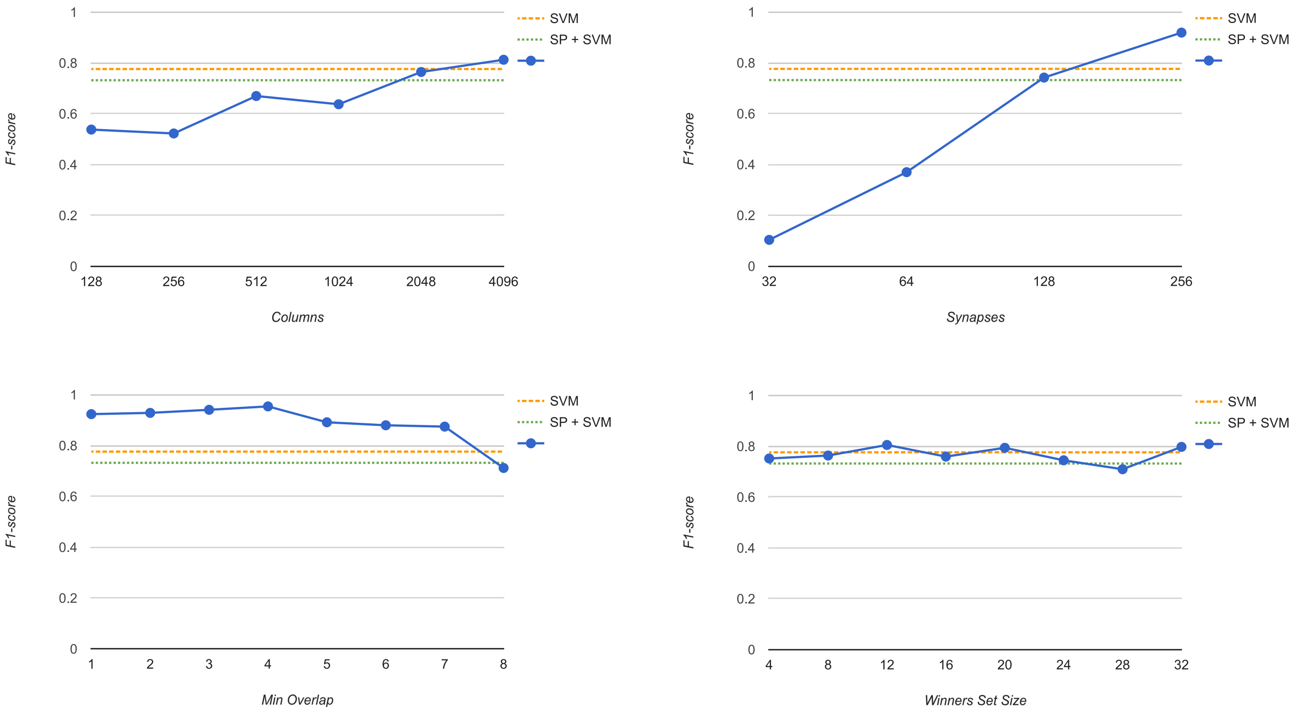


Fig. 9. Average F1 scores as a function of different SP configuration parameters

where the maximum is taken over all clusters and n is the number of all objects. The F1 score value ranges from 0 to 1, with a higher value indicating a higher clustering quality.

In each experiment presented in Fig. 9 one of the parameters was changed. 'SP + SVM' refers to the baseline results obtained with the proposed system using configuration values from Tab. II. It is worth noting that despite the superiority of the baseline 'SVM' setup, the 'SP + SVM' performance in selected cases is better than it is for 'SVM'. Especially, the number of synapses and the $min\_overlap$ value affects the performance of the module i.e. a rise in the number of synapses and a drop in the $min\_overlap$ value leads to better classification results. For every value of $winners\_set\_size$ the results remain on the same level with low fluctuation around the baseline. This results from the relationship between the inhibition radius and the $winners\_set\_size$ parameter. Change of the $winners\_set\_size$ is compensated by appropriate adaptation of the inhibition radius [4].

D. Acceleration results

A series of comparative tests were carried out for columns, synapses, $min\_overlap$ and $winners\_set\_size$. Two different test types were conducted, namely *GPU vs CPU OCL* denoted also as *OCL* and *GPU vs CPU kernel* referred to as *kernel* in the text. The first one accounts for the complete execution time of the examined procedures i.e. data preparation, data transfer in both directions and kernel execution [19]. The second test type embraces only kernel execution.

It should be noted that the GPU supersedes OpenCL CPU inhibition implementation and the discrepancy increases with increasing column numbers as it was presented in Fig. 10. Furthermore, OpenCL kernel performance is substantially better than its CPU counterpart (Fig. 11). However, when kernel launching procedures and data transfer are taken into account the speed-up is reduced. It is worth noting that it levels off at about 130x and 2.5x for kernel and OCL tests, respectively.

Fig. 12 and 13 show a change of speed-up as a function of the number of synapses connected to each column of a Spatial Pooler. The more synapses are connected, the greater the acceleration that is achieved. This results from the internal architecture of the overlap module (Fig. 6) which is, in essence, a hardware reduction operation performed within each GPU block. Fig. 13 depicts that both learning and testing phases of SP yield the same speed-up results. It is worth noting that, depending on the accelerator, there is a constraint on a maximum size of a work group, which directly translates to a limit in the number of synapses that can be accommodated by a single GPU block.

$Min\_overlap$ has a slight impact on performance and speed–up of the object classification system (Fig. 14 and 15). GPU execution time is gradually reduced reduced with a rise of $min\_overlap$. This results from the kernel implementation which allows for bypassing inhibition computation whenever overlap is lower than $min\_overlap$. For higher overlap values the number of zeros rapidly grows which leads to the rise of CPU/GPU speed-up.

$Winners\_set\_size$ is the number of 'winning' (having the highest overlap score) columns among the given column competitors in a contest to be chosen as active [4]. The number of neighboring columns which are taken into account impacts the computational effort since the columns are compared with all others within the inhibition range. Since $winners\_set\_size$ affects the inhibition radius, the larger the $winners\_set\_size$ is, the bigger the discrepancy in computation time between CPU and GPU, which is depicted in Fig. 17. Winners set computation may be perceived as a specific kind of reduction operation.

Fig. 18 presents the contribution of overlap computations to the complete inhibition execution routine. It ranges between 50 % and 75 % of total inhibition kernel calculation time.

It is worth emphasizing that overall OCL test results depend on data transfer, which in turn is related to data representation. Therefore, changing from integer to boolean data type will result in approximately 32–fold reduction of the amount of data to be transferred to the accelerator. Such a transition is unfortunately not available for all the data which are sent to the device, for instance $boost$ is of a float type and can not be easily mapped to boolean.

According to the authors' knowledge, it is hard to find papers which directly correspond to the research conducted in this work. Nevertheless, the following papers were examined: [20]–[22] which present results of video classification using UCF-101 dataset. The best systems presented in those papers are based on various architectures of Convolutional Neural Networks (CNNs) and achieve accuracy of 80% or more. It is worth emphasizing that despite similar performance in terms of the quality results, presented test setup is different mostly in terms of the dataset used for the experiments.

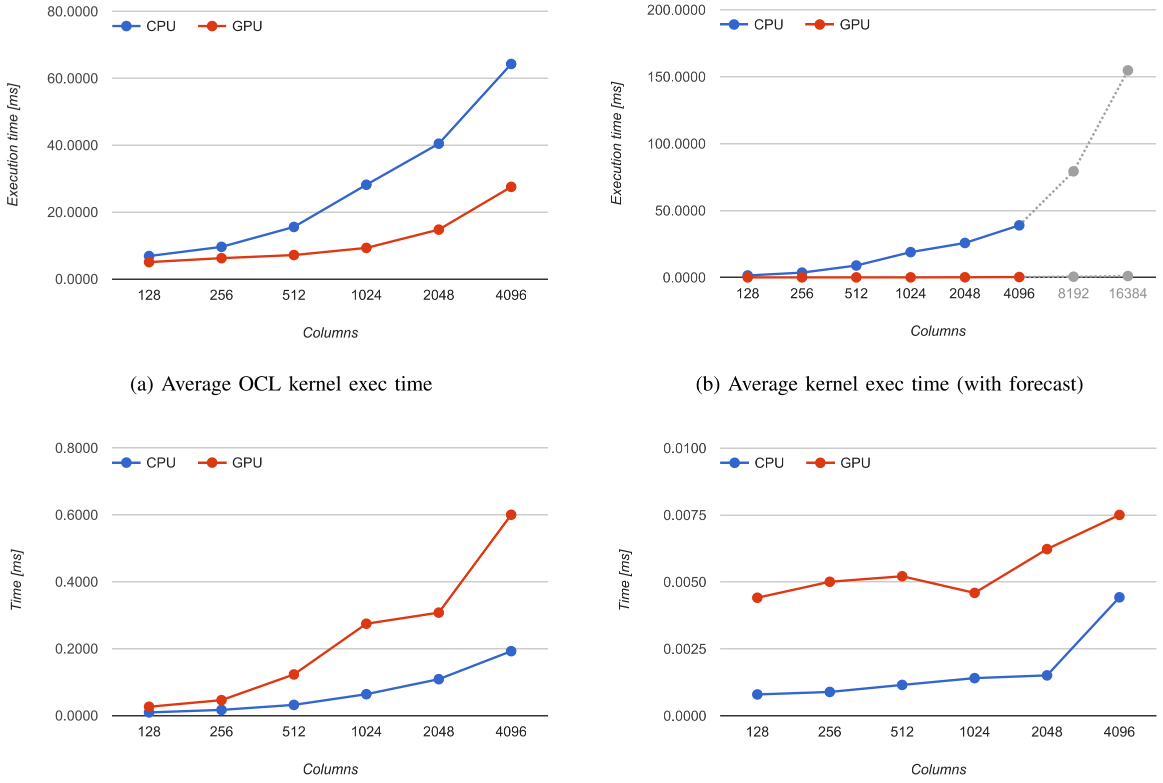
V. Conclusions and Future Work

This paper presents experimental results of using an HTM–based system for object classification in video streams. The classification accuracy of the system was examined through a series of experiments and the performance was given in terms of an F1 score as a function of the number of columns, synapses, $min\_overlap$ and $winners\_set\_size$. The system achieves the highest F1-score of 0.95 and 0.91 for $min\_overlap = 4$ and 256 synapses, respectively. A series of experiments with different hardware setups have also been conduced and CPU/GPU acceleration measured. The best kernel speed-up of 632x and 207x was reached for 256 synapses and 1024 columns. However, overall acceleration including transfer time was significantly lower and amounted to 6.5x and 3.2x for the same setup.

In future work, the authors are going to modify the preprocessing stage of the video processing flow and introduce TP. The authors are going to implement the most computationally–exhaustive routines in OpenCL and deploy the system on platforms equipped with GPU– or FPGA–based acceleration. This will enable conduction of experiments using video with a lower image reduction ratio and larger datasets as well as stacking several layers of SP.

Acknowledgment

I would like to thank my wife Urszula Wielgosz for her huge contribution to the preparation of the paper.



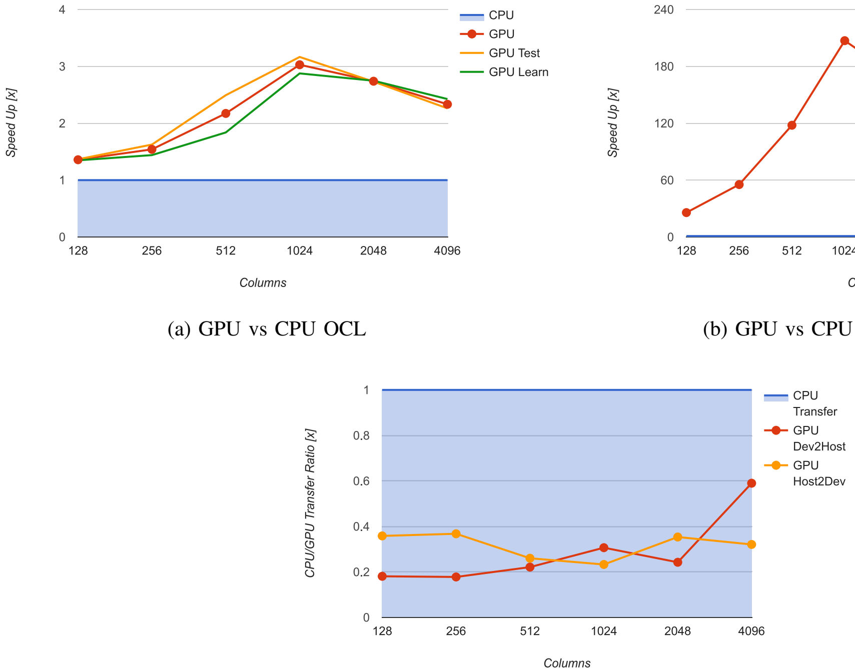
(a) Average OCL kernel exec time

(b) Average kernel exec time (with forecast)

(c) Average host–to–device data transfer time

(d) Average device–to–host data transfer time

Fig. 10. Profiling results for columns

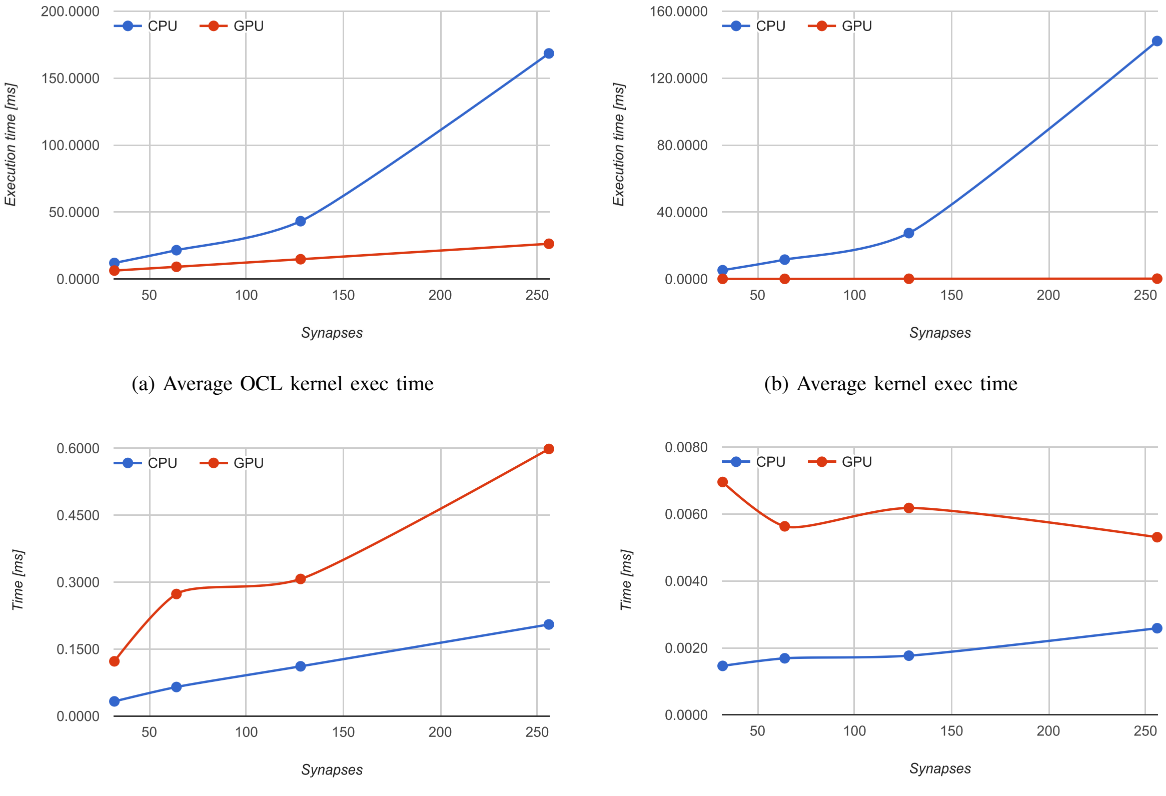


(a) GPU vs CPU OCL

(b) GPU vs CPU kernel (with forecast)

(c) GPU vs CPU data transfer

Fig. 11. Profiling results for columns



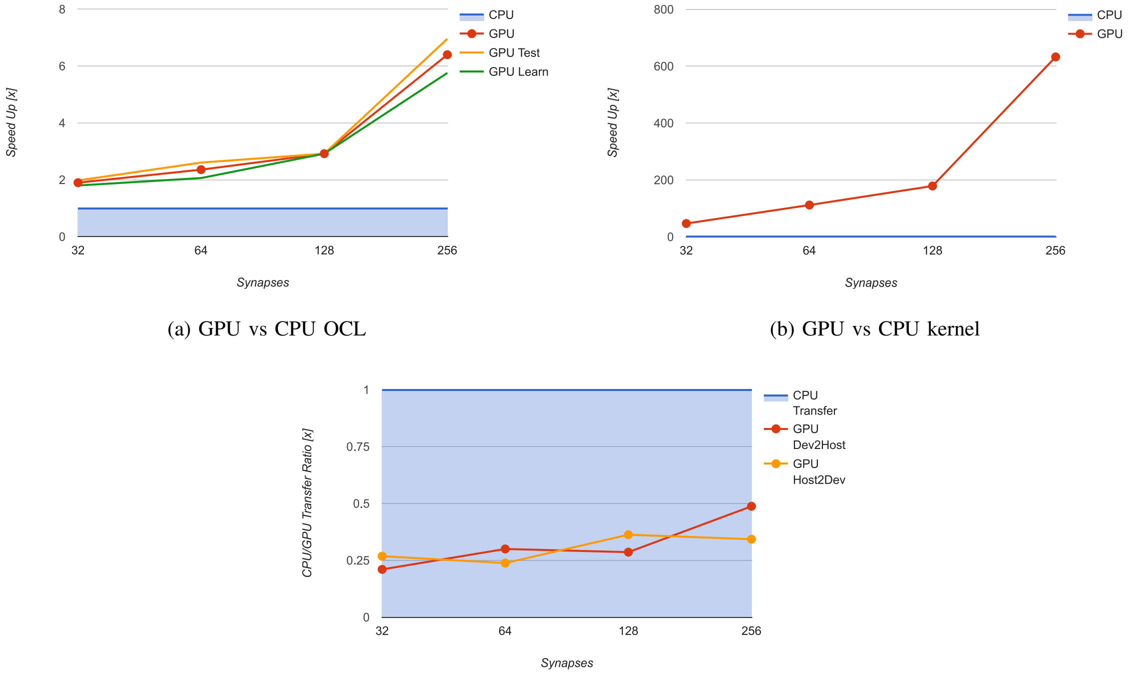
(a) Average OCL kernel exec time

(b) Average kernel exec time

(c) Average host–to–device data transfer time

(d) Average device–to–host data transfer time

Fig. 12. Profiling results for synapses

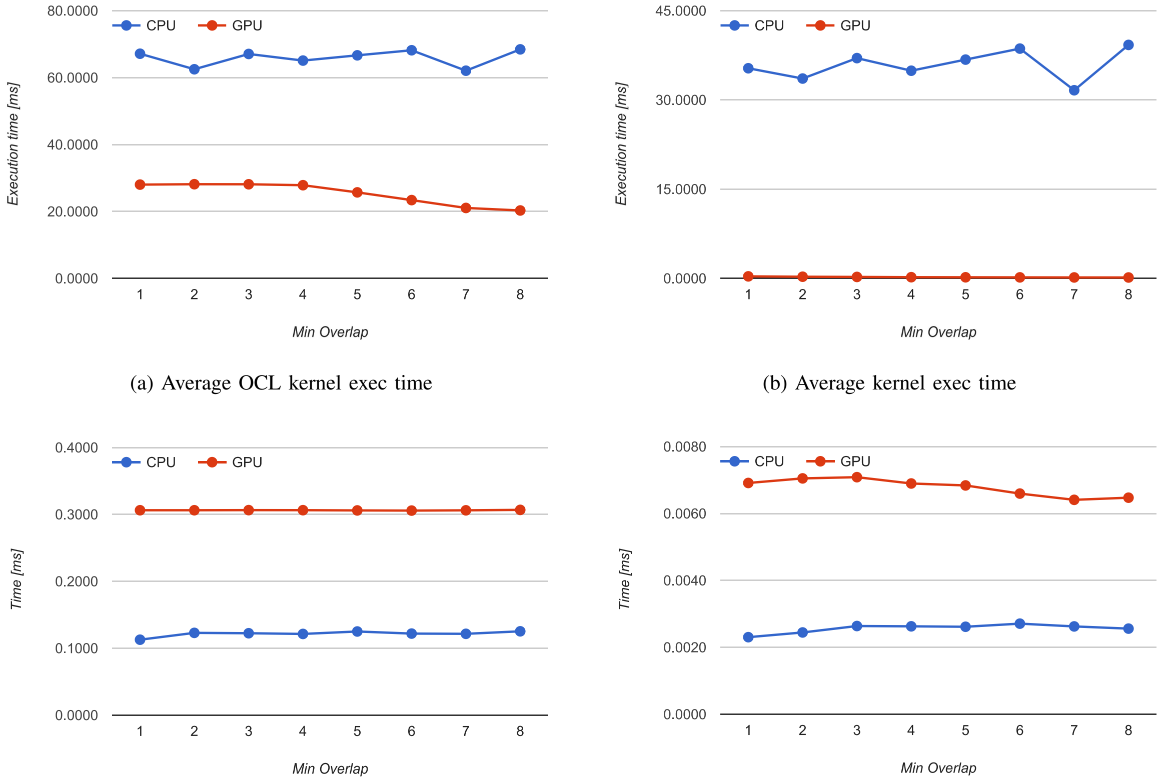


(a) GPU vs CPU OCL

(b) GPU vs CPU kernel

(c) GPU vs CPU data transfer

Fig. 13. Profiling results for synapses



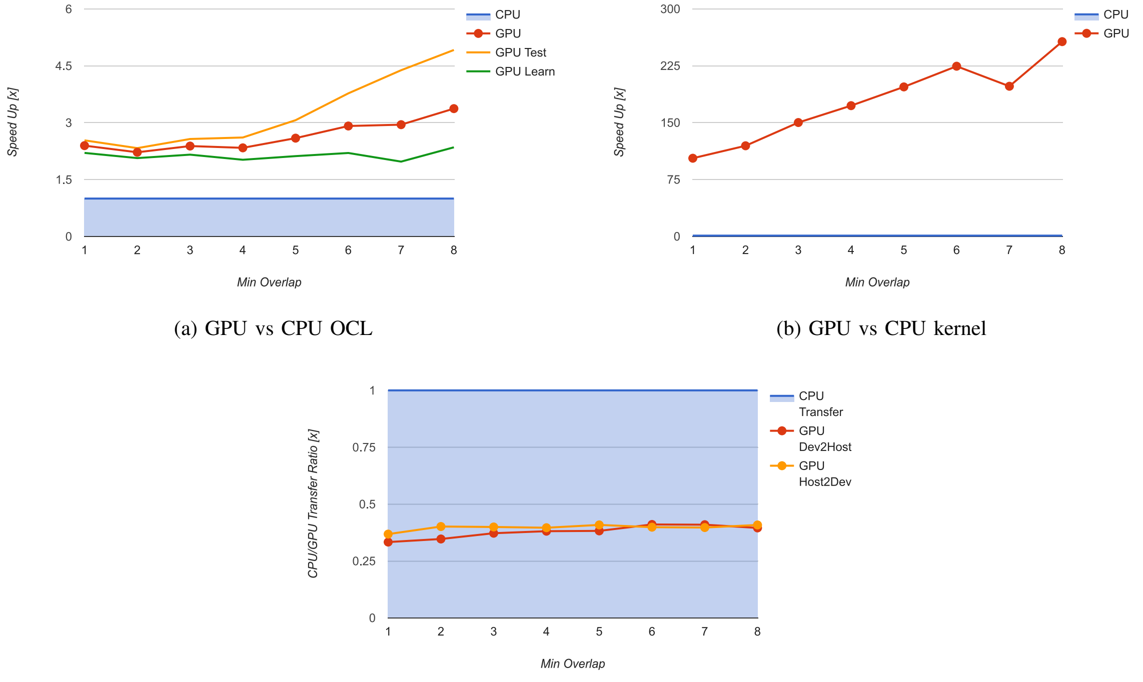
(a) Average OCL kernel exec time

(b) Average kernel exec time

(c) Average host–to–device data transfer time

(d) Average device–to–host data transfer time

Fig. 14. Profiling results for min overlap

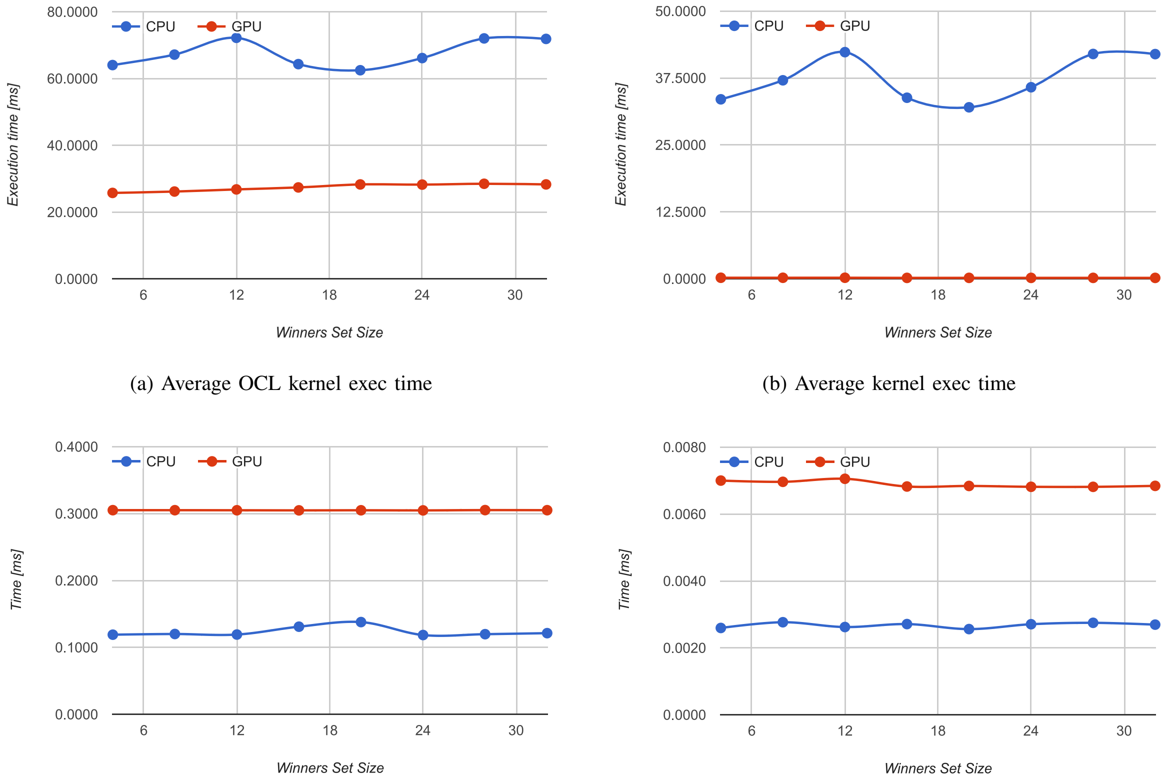


(a) GPU vs CPU OCL

(b) GPU vs CPU kernel

(c) GPU vs CPU data transfer

Fig. 15. Profiling results for min overlap



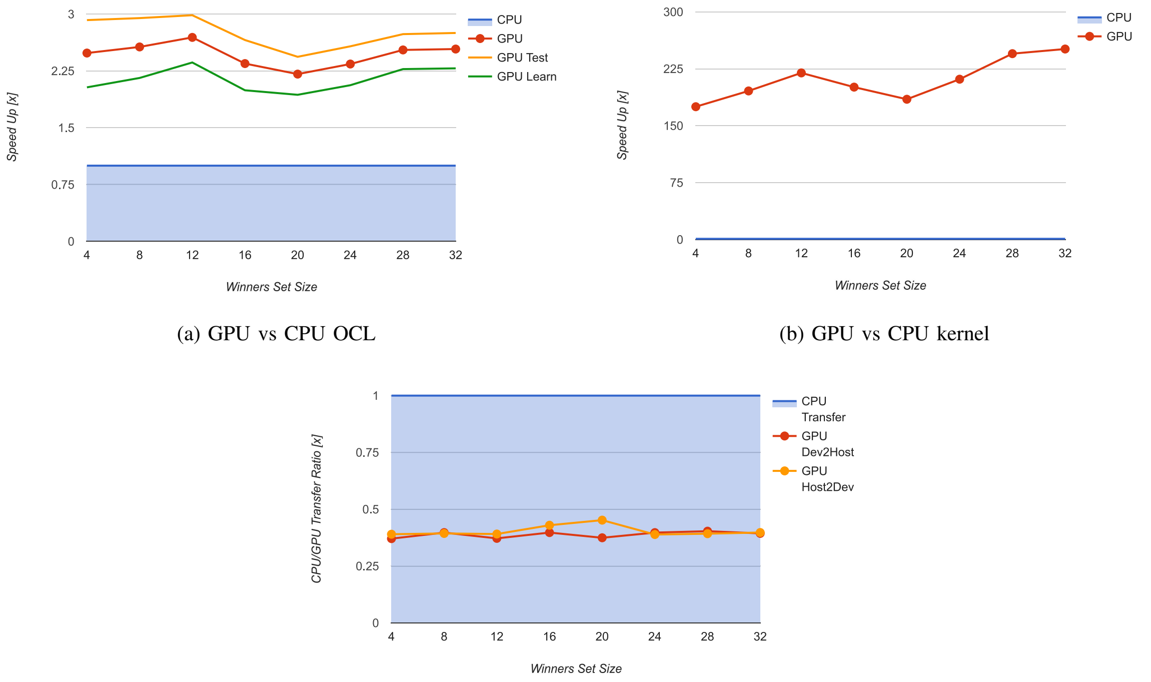
(a) Average OCL kernel exec time

(b) Average kernel exec time

(c) Average host–to–device data transfer time

(d) Average device–to–host data transfer time

Fig. 16. Profiling results for winners set size



(a) GPU vs CPU OCL

(b) GPU vs CPU kernel

(c) GPU vs CPU data transfer

Fig. 17. Profiling results for winners set size

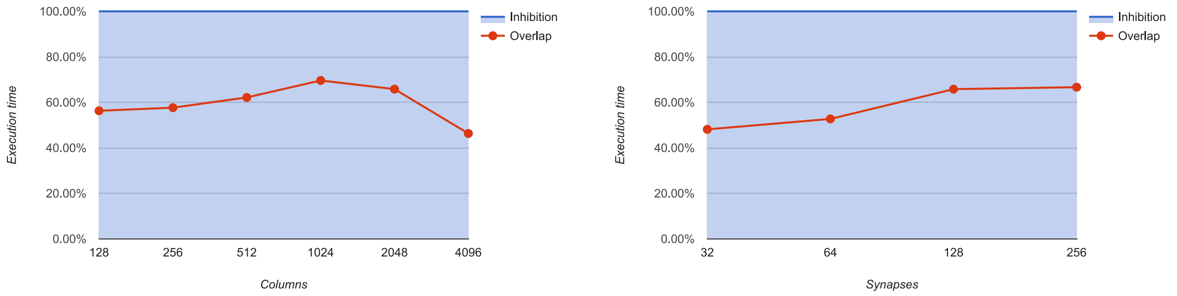


Fig. 18. Percentage of Overlap kernel execution time in whole Inhibition kernel execution time (on GPU)

References

[1] V. Mountcastle, "The columnar organization of the neocortex," *Brain*, vol. 120, no. 4, pp. 701–722, apr 1997.

[2] "The Human Brain Project - Human Brain Project," https://www.humanbrainproject.eu, (Accessed on 10.04.2016). [Online]. Available: https://www.humanbrainproject.eu

[3] "Custom Hierarchical Temporal Memory implementation," https://bitbucket.org/maciekwielgosz/htm-hardware-architecture, (Accessed on 12.04.2016). [Online]. Available: https://bitbucket.org/maciekwielgosz/htm-hardware-architecture

[4] J. Hawkins, S. Ahmad, and D. Dubinsky, "Hierarchical temporal memory including HTM cortical learning algorithms," Numenta, Inc, Tech. Rep., sep 2011. [Online]. Available: http://numenta.org/resources/HTM_CorticalLearningAlgorithms.pdf

[5] X. Chen, W. Wang, and W. Li, "An overview of Hierarchical Temporal Memory: A new neocortex algorithm," in *Modelling, Identification & Control (ICMIC), 2012 Proceedings of International Conference on*. Wuhan, China: IEEE, 2012, pp. 1004–1010.

[6] D. Rachkovskij, "Representation and processing of structures with binary sparse distributed codes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 2, pp. 261–276, 2001.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, sep 2009.

[8] P. Zhang, X. Zhu, and L. Guo, "Mining Data Streams with Labeled and Unlabeled Training Examples," in *2009 Ninth IEEE International Conference on Data Mining*, IEEE. Miami, USA: IEEE, dec 2009, pp. 627–636.

[9] X. Lu, C. Zhang, and X. Yang, "Online video object classification using fast similarity network fusion," in *2014 IEEE Visual Communications and Image Processing Conference*, IEEE. Valletta, Malta: IEEE, dec 2014, pp. 346–349.

[10] R. N. Hota, V. Venkoparao, and A. Rajagopal, "Shape Based Object Classification for Automated Video Surveillance with Feature Selection," in *10th International Conference on Information Technology (ICIT 2007)*, IEEE. Rourkela, India: IEEE, dec 2007, pp. 97–99.

[11] M. K. Islam, F. Jahan, J.-H. Min, and J.-H. Baek, "Object classification based on visual and extended features for video surveillance application," in *Control Conference (ASCC), 2011 8th Asian*. Kaohsiung, Taiwan: IEEE, 2011, pp. 1398–1401.

[12] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, vol. 18, no. 2, pp. 130–140, apr 2007.

[13] P. Devarakota, M. Castillo-Franco, R. Ginhoux, B. Mirbach, S. Kater, and B. Ottersten, "3-D-Skeleton-Based Head Detection and Tracking Using Range Images," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 8, pp. 4064–4077, oct 2009.

[14] F. N. Khan and S. A. Khan, "Real-time object based single-stream to multi-stream network enabled multimedia system using an adderless reconfigurable fast area correlator processor," in *8th International Multitopic Conference, 2004. Proceedings of INMIC 2004.*, IEEE. Lahore, Pakistan: IEEE, 2004, pp. 688–693.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, aug 2013.

[16] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[17] "Blender project - Free and Open 3D Creation Software," https://www.blender.org/, (Accessed on 12.04.2016). [Online]. Available: https://www.blender.org/

[18] "HTM Test Datasets," http://data.wielgosz.info, (Accessed on 02.07.2016). [Online]. Available: http://data.wielgosz.info

[19] A. Klöckner, N. Pinto, Y. Lee, B. Catanzaro, P. Ivanov, and A. Fasih, "PyCUDA and PyOpenCL: A Scripting-Based Approach to GPU Run-Time Code Generation," *Parallel Computing*, vol. 38, no. 3, pp. 157–174, 2012.

[20] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4694–4702, jun 2015.

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014, pp. 1725–1732.

[22] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting Image-trained CNN Architectures for Unconstrained Video Classification," *ArXiv e-prints*, mar 2015. [Online]. Available: http://arxiv.org/abs/1503.04144

Performance Analysis of Proposed Congestion Avoiding Protocol for IEEE 802.11s

Kishwer Abdul Khaliq[1]
Department of Production Engineering,
IGS, Universität Bremen,
Germany

Amir Qayyum
[1]CoReNeT,
Capital University of Science and
Technology (CUST), Islamabad,
Pakistan

Jürgen Pannek
Department of Production Engineering
Universität Bremen, BIBA Bremer Institut
für Produktion und Logistik GmbH,
Germany

***Abstract*—The wireless technology is one of the core components of mobile applications with mobility support at low deployment costs. Among these, Wireless Mesh Network (WMN) is one of the technologies that supports mobile users for un-disrupted, reliable data connectivity, provides high bandwidth even in areas, where access of such services is difficult. Additionally, it features capabilities like self-configuring, self-healing, and self-organizing. IEEE proposed a MAC standard for WMN enhancements named IEEE 802.11s for multi-hop networks. Within this standard, the mandatory routing protocol called Hybrid Wireless Mesh Protocol (HWMP) is proposed for efficient utilization of resources to achieve high bandwidth at MAC layer. To improve this protocol, a congestion avoiding protocol was proposed, which utilizes alternate paths just before the congestion state is reached. The proposed technique does not add any overhead, it utilizes congestion notification frame, which is already part of standard. This paper discusses simulation results of the proposed routing protocol against the existing HWMP protocol for packet delivery fraction, throughput and delay. The results indicate that the proposed technique significantly improves performance of IEEE 802.11s.**

***Keywords*—Wireless Mesh Network; IEEE802.11s; Congestion Control; Congestion Avoidance; Routing Protocol; HWMP**

I. Introduction

The wide variety of interesting applications including broadband home networking, collaborative networks, building automation system, enterprise networking are using wireless technologies. These technologies are popular because of their attractive offered advantages, e.g. support for mobility and easy deployment. These are also used in number of strategic and smart health-care applications [1], commercial applications [2], automation [3] and disaster management [4]. Apart from the information and specific applications, currently users have more interest in use of general purpose applications like infotainment, online video gaming and streaming applications [5]. With the use of multimedia and interactive sessions, applications require high bandwidth and internet connectivity without latency. A Multi-hop wireless network is the best choice to fulfill user's need because it has the advantages of both ethos multi-hop and broadband access. This multi-hop wireless network is named Wireless Mesh Network (WMN), and it offers ease of deployment in rural and hilly areas to provide broadband services. The most efficient applications supported by WMN are broadband wireless access, industrial and business applications, smart health care, transportation management systems, production, hospitality, warehouses and provisional venues [6]–[10]. WMN is a special type of adhoc network with self-healing, self-configuring and self-organizing capabilities, and is also used for deploying wide variety of applications like e-applications [11], public safety and crises management applications [12], building automation control [13], emergency and safety applications [14]–[16].

In WMN, the mesh nodes have capability of a relay station and these nodes can communicate directly without involvement of central entity. Many solutions have been provided by different organizations for WMN [17], where it has three types of nodes. These are mesh clients, mesh routers and gateways. Furthermore, this technology can operate in three modes, i.e. infrastructure, client, or Hybrid. The first mode provides backbone access to the conventional clients and integration with existing networks [18], the second works in adhoc mode and the third is combination of both. The two-tier IEEE 802.11s standard is proposed by IEEE for WMN [19], which includes backhaul and access tier [20].

In IEEE 802.11s [21], adhoc mode allows communication between nodes without any central entity (i.e. Access Point (AP)). WMN in Figure 1, gives an overview of WMN architecture in hybrid mode, basically it is a combination of Independent Basic Service Set (IBSS) and Extended Service Set (ESS). There are mainly three types of nodes, i.e Mesh point (MP), Mesh Portal (MPP) and Mesh Access Point (MAP). MP acts as a router excluding AP functionality, and use wireless links for connectivity to other nodes. Hence, internal mesh LAN is not an ESS. MAP has combined functionality of AP and MP. It gives association to the station. MPP includes MP and gateway functionality. MPP is responsible for handling entry and exit of MSDUs from WMN to other connected network. IEEE 802.11s MAC is the enhancement in the existing IEEE 802.11 MAC, but also have some additional functionality like routing protocol. This routing protocol is handled at MAC layer whereas layer 3 routing protocols are needed at MPP for path selection, while communicating with other networks [17]. The protocol stacks of the IEEE 802.11s on each type of type is shown in Figure 2.

The ongoing research is focusing on the functions of 802.11s MAC, which includes QoS (performing priority Control), congestion control and admission control. In addition to it, the enhanced MAC also includes a functions for achieving spatial frequency reuse, and to avoid performance degradation due to hidden and expose nodes problems [22]. To control congestion in network, many researchers consider its queue

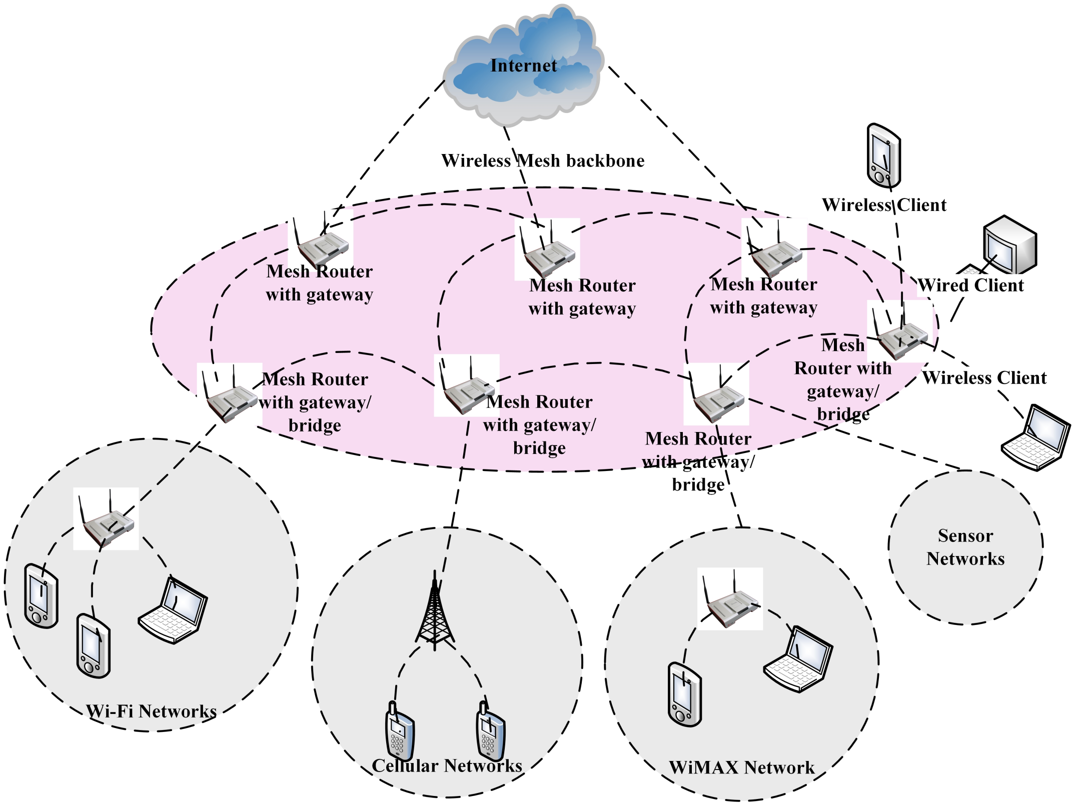


Fig. 1: Infrastructure/backbone WMN

mechanism, link capacity and routing protocols to get required results. Congestion is an important problem domain and it occurs when incoming network traffic load at a router is greater than the out going traffic rate [23]. In wireless scenario, it is predicted that packet lost or queue overflow may one of the reason for congestion. The reason behind this are the wireless communication issues which includes greater error rate due to wireless channel, wireless bandwidth etc. and the shared characteristics of the wireless channel. When packets are delayed due to mentioned issues, the increase in buffered packets in queue leads to congestion. In this regard, IEEE 802.11s does not specified any congestion control mechanism. In literature, researcher proposed a number of congestion control mechanisms to resolve this problem, but every protocol has own limitations. In [24], we proposed a routing protocol to avoid congestion mechanism to address this problem which increase the throughput of the network. In this paper, we evaluate the proposed technique for congested scenarios and found that this protocol helps to reduce delay and packet loss. The remaining organization of the paper is as follows; Section II includes literature survey in details for congestion control protocols, Section III includes the proposed technique and Section IV includes the simulation and result analysis. Last Section V concludes our research work and discusses future work.

II. Literature Review

All types of networks are facing Congestion issue, when they have to handle data more than the available capacity. It occurs when the incoming number of packets and outgoing number of packets are greater than the available network capacity. Network capacity is measured in term of network resources, available bandwidth or buffer of the network [26]. In wireless network, when one node has data to send, it accesses wireless channel and in this mechanism one node can transmit data at one time to share characteristics of it. Sometime, the increase in wait time to access wireless channel also increase packet delay and resulting queue length leads to congestion. As it is known that traffic is aggregated at the portal nodes, therefore, in a case of greater traffic load, MPs on the outer edges of the network suffer low throughput and greater packet loss in absence a congestion control mechanism [23]. An optional congestion control mechanism names hop-by-hop congestion control mechanism is outlined in the standard draft [21]. Each MP in hop-to-hop monitors congestion level by monitoring the incoming and outgoing traffic in its buffer. When the traffic load reaches at congestion specified threshold, the congested node should notify to its neighboring nodes to control their traffic. This mechanism includes three basic steps. The first step involves monitoring processes, in which each MP in network has to monitor its queue level for congestion rate or minimize the queue size by regulating the data traffic rate. The second step involves the notification on detection of congestion. The congested node broadcast Congestion Control Notification Frame (CCNF) to immediate nodes on its detection. The third step involves a control process. The nodes who receive CCNF, they limits their traffic rate according to service differentiation criteria. In addition to it, the CCNF also contains expiry information for

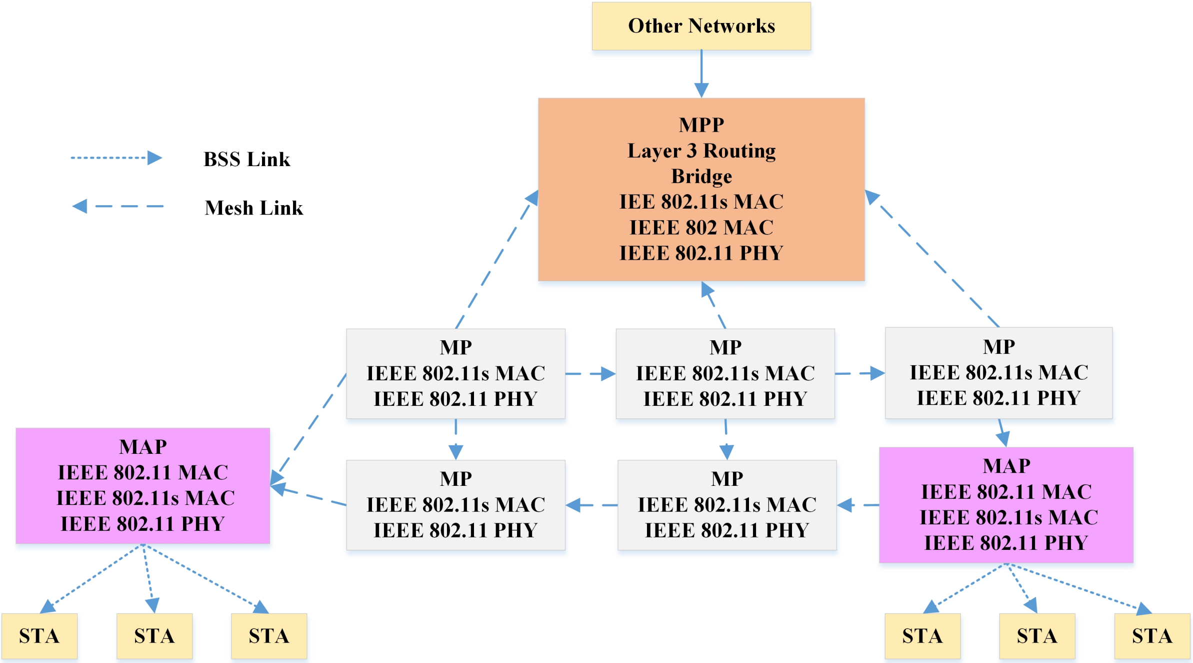


Fig. 2: Protocol Stack of IEEE 802.11s [25]

notification. In some cases, channel rate is also considered to restrict data-rate. Hence, a mesh node can also use this criteria for controlling data-rate [27].

In [28], authors proposed a modification in hop-by-hop mechanism by including feedback mechanism in distributed manner. This technique requires two NICs on each node at the same time, which operates independently. The mechanism is first derived algorithm for end-to-end, then it is further derived for hop-by-hop congestion control to control source rate end-to-end. This algorithm works with the assumption of total knowledge of each flow on on each intermediate node. The controlling algorithm is responsible for monitoring incoming and outgoing transmissions and it performs computation on each relay node to sum all congestion states and maximum transmission rate for each flow. The intelligent part of it is that it selects smaller value for maximum possible rate for transmission. The drawback of this mechanism is additional cost for extra NICs and increase in overhead due to continuous feed back mechanism. There is an additional processing and synchronization cost because of combined algorithms to control congestion on each node.

In [29], the authors proposed another algorithm to provide end-to-end max-min-fairness to each flow. This co-ordinated congestion control algorithm is designed to deal with inters and intra-flows using multi-hop wireless links. On each wireless link, max-min-fairshare is computed continuously on assigned bandwidth and each flow uses allocated share in a fair way. In the whole mechanism a gateway is a central coordinator, which is used for traffic engineering. Similar type of mechanism is proposed in [30]. These both algorithms solve the issue of unfair channel sharing. However, these algorithms do not help in congested scenarios, because mechanism does not provide any feedback mechanism to limit traffic. In [31], the authors proposed a source based congestion control algorithm called WCP for multi-hop WMN, where source node is maintaining transmission information for each flow. It uses Additive increase and multiplicative decrease (AIMD) for controlling transmission rate. The algorithm uses WCPCap to estimate capacity of its neighbors and share among contented nodes. But the problem with this algorithm is that it is not providing any solution if a node receives and forward data for multiple nodes. Because of relay functionality and multiple flow maintenance on each node, the delay increases. Moreover, battery conceptions is greater due to additional computation for each flow on each node as compared to simple mechanism.

In WMN, there are two common type of congestions i.e. intra and inter-mesh congestion. Multiple algorithms are designed to resolve intra-mesh congestion. These algorithm also use congestion notification to control it. In [26] two algorithms i.e. Total Congestion Control (TCC) and Link Selective Congestion Control (LSCC) are proposed for intra-mesh congestion, but these algorithm do not provide very efficient solution. In TCC, CCNF is sent in local vicinity when congestion state reaches. The immediate nodes on receiving notification, block all traffic. In LSCC algorithm, on receiving congestion notification, the immediate nodes limit the traffic for specific link by blocking the data packets for a specific destination. An expiry time period is also included in CCNF, and flow resumed on expiry of notification. CCNF also contained information about the congested link, and when a node receives this frame, it blocks traffic only for the mentioned link. In [32], another algorithm is proposed which blocked traffic selectively. This algorithm is known as Path Selective Congestion Control (PSCC). This algorithm blocks traffic only for specified destination when a node receives notification on congestion occurrence. The CCNF frame includes information about specified flow. For the announcement of specific destination, this algorithm requires modification in the standard CCNF. Furthermore, on receiving modified CCNF, a node only blocks sending data for a specific destination, but it continues receiving for specified node. The scenario becomes more complicated when CCNF frame is further broadcast to immediate nodes in a continuous chain. These algorithms resolve congestion problem in few scenarios of multi-hop WMN.

Consider Figure 3 (a) for a congested scenario, it shows a congested link between mesh node *C* and mesh node *D*. A queue size is monitored at node *C*, and when it is reached at the specified value, the node broadcasts the CCNF to the immediate nodes to limit traffic for node *C*. In the current scenario, node *E* and node *B* are in the neighbors. When these node receive notification, they stop transmitting data for node *B* until the expiry of notification. In the mention scenario when we apply TCC algorithm, the immediate neighboring nodes stop data transmission but they continue reception from own neighbors. These nodes buffered all received data instead of forwarding till expiry of notification. If congested link could not resume from congestion, followed by another notification, then this delay cause congestion on neighboring nodes because of queued data. When these nodes also reach to a congestion state, and they also broadcast CCNF. If process continues, the whole network becomes congested.

In the same scenario when we apply LSCC, the immediate neighboring nodes stop data transmission for node *C* until the CCNF time expired on receiving notification. During the notification expiry time, node *B* queues all the received packets for the node *C* only and forwards rest of the data traffic to other nodes. In this special case node *C* also continues its local and global traffic transmission to its neighbors , which are node *E* and node *B* in this scenario. When the buffered data in the queue of the node *B* reaches to the specified threshold, the node *B* also broadcasts CCNF to its neighbors in vicinity. In the considered scenario, there neighbors are node *A*, node *F* and node *C*. These nodes on receiving notification, block the traffic for node *B*. But the node *C* whose traffic is already blocked due to last notification, will not receive this notification. In absence of any notification, the node *C* as depicted in Figure 3 (b) continues its local data transmission or if already queued global traffic to node *B*, which results the packet lost. This situation becomes worst if node *A* also becomes congested and broadcasts CCNF. As MPP is responsible for in-going and out-going data in WMN, therefore MPs and MPP have the most congestion chances, specially for the case of bandwidth hungry applications. When it is occurred, packets overflow from buffer regardless of the number of hops the packets already have been traversed.

In this problem domain, authors in [24] proposed a technique called Congestion Avoidance Hybrid Wireless Mesh Protocol (CA-HWMP), which gives preference to avoids the congestion before its occurrence to improve packet delivery and improve network throughput. In this paper, we included the details about the proposed idea, algorithm and its limitations in Section III. Then we performed number of simulations to evaluate the behavior of CA-HWMP for different application data-rate for network throughput, packet delivery fraction and most important the end-to-end delay dealing with congestion issue. Section IV discusses all simulation results. Section V concludes all works with future directions.

III. Proposed Mechanism

Congestion Control mechanism works when congestion is already introduced in the network. The proposed mechanism focused on prevention of it. In WMN, routing is performed at the Data Link layer, and the proposed mechanism utilizes this routing protocol for congestion avoidance with small modification in basic mechanism. That is why it is named as Congestion Avoidance Hybrid Wireless Mesh protocol (CA-HWMP) [24] and it uses HWMP in IEEE 802.11s with modification in mechanism for congestion avoidance. It includes three steps, monitoring, notification and intelligent re-routing. In first step, it monitors queue length on each node for each flow. The second step is to notify neighbors when it reaches to specified level. In the thirs step, the neighboring node calculate alternate path for destination by consider queue level to avoid congestion again. The proposed mechanism did not change the basic four information elements i.e. Path Request (PREQ), Path Reply (PREP), Path Error (PERR), and Root Announcement (RANN).

Consider a scenario depicted in Figure 4, it includes nine nodes in total. This scenario includes *G* as a source node and *C* as a destination node. Furthermore, node *A*, *F* and *C* are in the neighbor of node *B*. Node *G* and *H* are in the neighborhood of *A*. Node *G* sends data to node *C*, the optimal path selected by its routing protocol is $G - - > A - - > B - - > C$. In the selected best path, the immediate link to *C* for a source node *G* is node *B*. For a link *B–¿C*, node *B* monitors its queue length, when this queue length at node *B* reaches at specified threshold, it broadcasts the CCNF frame to its immediate neighbors. The notified nodes, who have flow for node *C* or through node *C* performs reactive mechanism to find path for the specified destination excluding congested link. In the given scenario, the immediate neighbors of node *A* are *G, H, F and B*, which receive PREQ. Node *C* and Node *G* discard PREQ request, because first is congested and second is itself a source node. The PREQ is forwarded until it reached to the destination node. The destination node reply for the path by sending unicast PREP to the source. The new route establish from source mesh node *G* to destination node in the given scenario is $G - - > A - - > F - - > I - - > C$. This procedure resumes data transmission via new calculated alternative link. This mechanism is not only good for the congestion avoidance but also reduce load-balancing at specific link/node in the multi-hop WMN. The queued data packet in the absence of congestion avoiding protocol, will now forward to destination node using this new established path. This mechanism reduces the packet lost, which was taking place due to queue overflow. This protocol allows the data transmission on the alternate route instead to wait the positive signal of congestion to restart transmission on existing track.

The proposed technique CA-HWMP which is basically a modification in the default protocol of IEEE 802.11s i.e HWMP. The Algorithm 1 works in active mode. Whenever a node have data to transmit, first it establishes a path for destination. For path selection, it broadcasts the PREQ to its immediate node. The receiving nodes forward the PREQ according to basic rules of the default protocol and additional to that they will check their queue level. If it is below to the

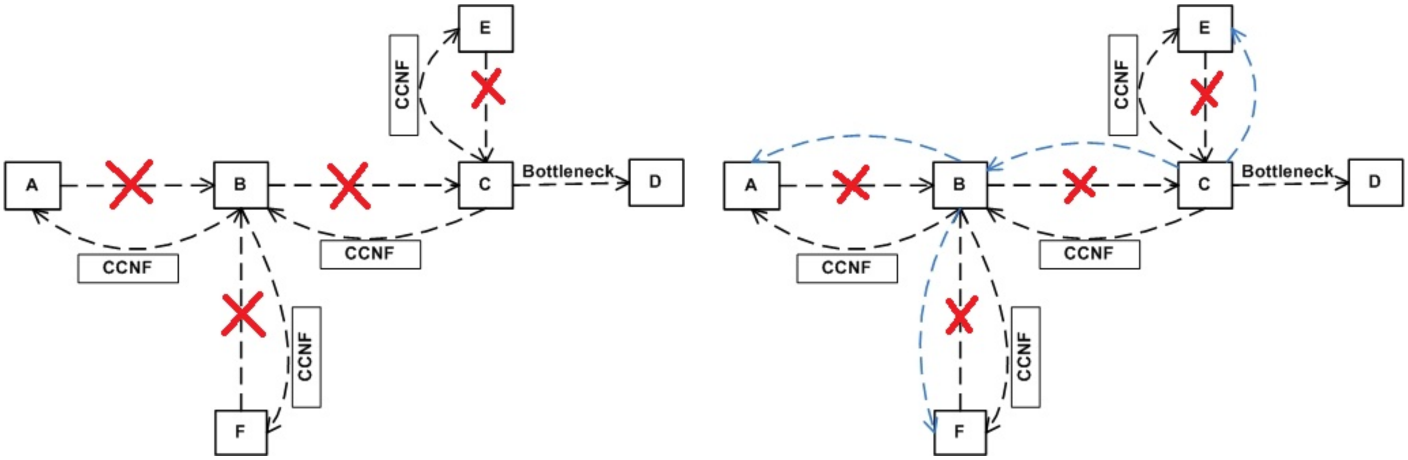


Fig. 3: (a) Congestion Scenario Using TCC; (b) Congestion Scenario Using LSCC;

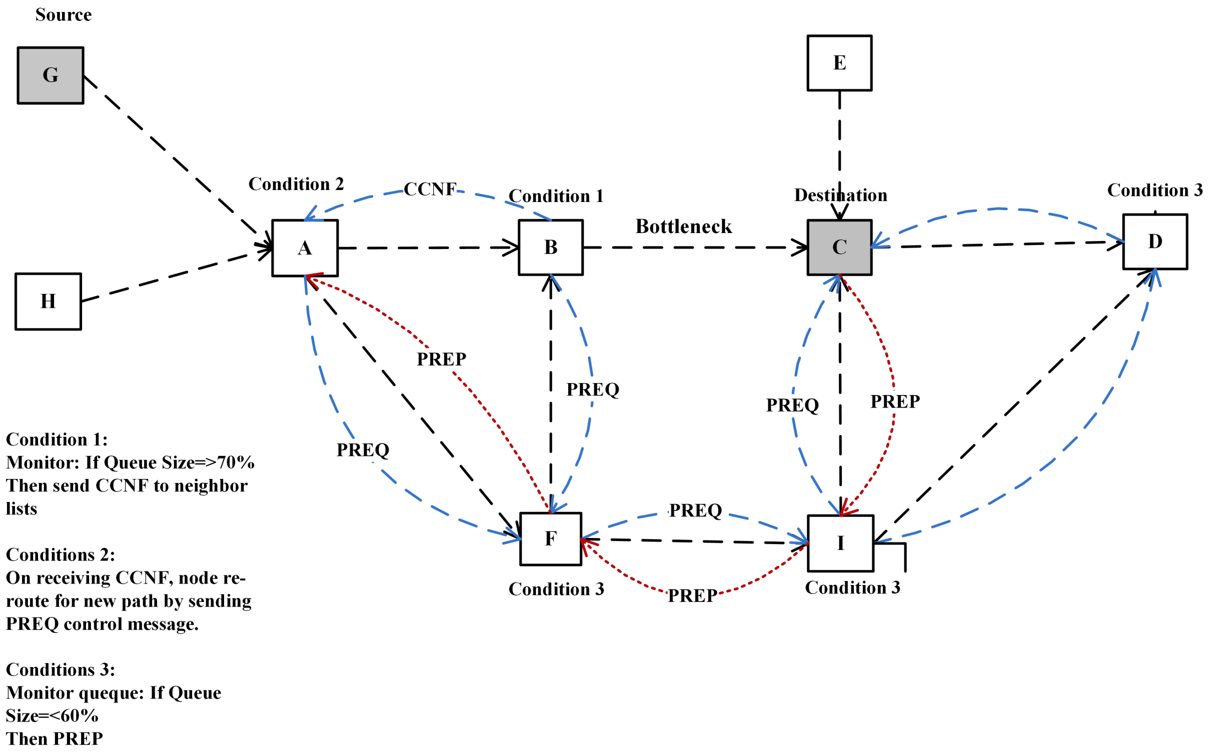


Fig. 4: CA-HWMP protocol Mechanism [24]

defined value, it forwards PREQ to other nodes. Finally, when the PREQ source node receives PREP, the path will establish.

Every mechanism has advantages and disadvantages. The proposed mechanism also has some limitations of CA-HWMP. Our proposed technique works well in a scenario, where we have possibility of alternate ways to re-route traffic. Although there are 80% chances of availability of alternate routes. Nonetheless, in the absence of alternate path, our proposed technique adapts standard available procedure. This protocol gives advantages of alternate routes. In absence of this technique, nodes received packets from neighboring nodes and queue them until CCNF expiry time reached. But in presence of it, utilization of alternate paths add benefits. It is doubted that the proposed protocol may have some scalability issue. It is the common practice, for route calculation few message exchanges between nodes in the network. If we have greater nodes in the network then this increase routing overhead. In a wireless network, with the increase of nodes in the network channel contention also increases and results in the increase in the wait time.

Algorithm 1 Algorithm for CA-HWMP

```
(Variables)
$SourceNodeData$ : Boolean variable for data status, Value1,0
$QueueMax$ : Maximum queue length
$PREQ$ : Path Request
$PREP$ : Path Reply
$Path$ : one hop path
$TO$ : Target-only flag
$RF$ : Reply-and-Forward
$SequenceNum$ :Sequence number for PREQ
$ownSequenceNum$ : temporary value use for sequence number saved at intermediate/ destination node
(Main Algorithm)
$If((SourceNodeData == true)||(QueueMax => 65\%))$
Broadcast PREQ
upon receiving PREQ$if(QueueMax =< 60\%)$
Discard PREQ message
$elseif(SequenceNum > ownSequenceNum)$
Update Path
$if($New Path created/ Modified$)$
forward PREQ
(Flags)
$TO = 1$ : Target-only sends PREP
$TO = 0andRF = 0$ : intermediate node sends uni-cast reply to source with Path, and does not forward PREQ
$TO = 0andRF = 1$ : The first intermediate node with the Path, sends reply to source. It also change TO=1 and forwards PREQ
```

TABLE I: Considered Parameters for Simulation

Operating System	Linux Distribution Fedora Core
NS-3 version	NS-3.14
Wifi Standard	IEEE 802.11s
Mobility Model	Constant Position Mobility Model
Number of Interface	1
RTS/CTS	Disable
Trace Module	Flow Monitor
Traffic Flows	Constant-bit rate (CBR)
Flows Varies (Kbps)	100, 150, 200, 250, 300, 350
Packet Size (KB)	1024
Transport Layer Protocol	UDP
Routing Protocols at MAC	HWMP, CA-HWMP
Number of Nodes in Grid	4, 9, 16, 25, 36, 49, 64
Transmission Range (m)	170
Simulation Time (Sec)	240

IV. Simulation and Result Analysis

The main objective of this paper is to evaluate and discuss results of proposed CA-HWMP protocol in different scenarios. NS3 is used for protocol implementation, which is an open source simulator. It provides support and implementation flexibility of module implementation for wireless mesh network. By taking advantages of it, we patched CA-HWMP successfully into the already available mesh module using C++, then we use scripts to evaluate protocol.

In Table I, the general simulation parameters are listed that we use in our selected scenarios. We perform simulation on the Linux Distribution Fedora Core using NS-3.14 version. It includes built-in supports of IEEE 802.11s and we used it as bench mark to compare with the implemented own module. To create congestion scenarios, we used On-off (CBR) application, which transmits data at a constant bit rate. During simulation scenario implementation, we focused on the queue level monitoring while increasing traffic rate slowly. Therefore, in our simulation scenarios, the used data-rate varies from 100Kbps to 350Kbps on UDP transport layer protocol. We started our simulator from 4 nodes and then we increases this number exponentially. Therefore we use grid topology for nodes positions. To observe closely, the number of nodes increases in both dimensions in each simulation scenario. The grid topology is represented in form of X and Y-axis as $m \times n$ where ”m” represents the number of nodes on X-axis and ”n” in Y-axis. The distance between two nodes is 170m. Multiple simulation scenarios have been considered to observe the effect of application data-rate on throughput, packet delivery fraction (PDF), and end-to-end delay. The considered evaluating parameters are most effected in congestion scenarios. We have selected multiple simulation scenarios, and in each scenario the possibility of multiple path is varied because the simulation is performed on varied number of nodes. The chosen grid topology ($m \times n$) consisted of MP nodes. Figure 5 represents this topology where ”m” indicates the number of the nodes in the X-axis and n indicates the number of nodes in Y-axis. The first simulation run uses 2×2 grid then the increase in values of ”m” and n was additive.

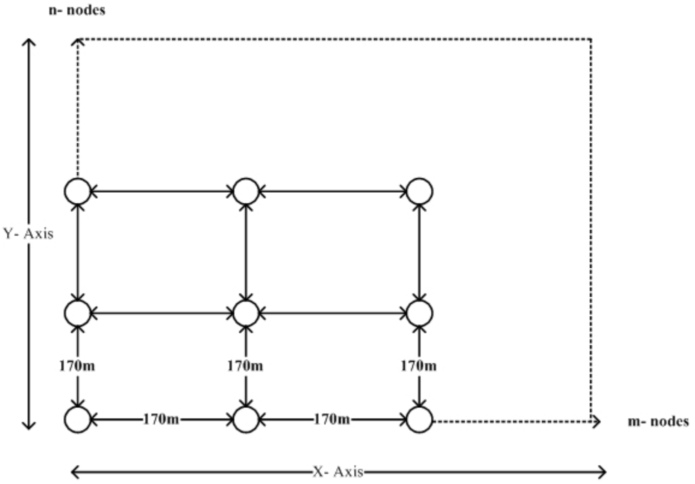


Fig. 5: $m \times n$ Grid Topology

A. Effect of Application Data-rate on Throughput

Throughput is one of the evaluating parameter in network simulations. To analyze the network behavior for this parameter, we fixed the traffic generating node. As the network contains different number of nodes along application data-rate variation, therefore we fixed the traffic generating nodes upto 50%. To make scenario more realistic, we choose the source and destination nodes at run time. The nodes participate in path selection active and those node who have data to send use path selection. The reason behind to limit traffic flows up to 50% is that the utilization of alternative paths can be observed correctly. In this scenario, the used application is CBR where its traffic varies from 100Kbps to 350Kbps, and the considered nodes varies from 4 to 64. The maximum data-rate is 230Kbps, therefore the device maximum data-rate is also fixed 350Kbps. The computed value is the average throughput of the network. We generated graphs on computed value and graphs in Figure 6 (a) to (f) represent the network throughput on varied node density and data-rates. The variation in nodes in the grid topology is shown on X-axis where grid varies from 2×2 to 8×8, while resultant throughput is shown on Y-axis.

The Graph in Figure 6 (a) shows network throughput for 100Kbps application data-rate, and the node density was from sparse mode to dense. Consider first case when the grid has 4 nodes, each node is in the direct access of one another and can transmit data directly with hop count zero. Then this grid size is increased from 2×2 to 3×3, the performance of both protocols were again similar. The reason is simple, in 9 nodes grid, node-1 sends data to node-9 in the network; the intermediate node relays to the destination node. In this case, the best path has maximum one hop. The device has a capability to transmit is more than 3 times greater than the application data-rate. The only case is that, if more than 3 nodes transmit data using one relay node then the relay nodes queued the excessive packets. When this queue becomes full, it leads to congestion. We have very less nodes in the network, hence, are less chances of congestion on relay nodes. However, throughput of CA-HWMP negligible better than HWMP due to use of alternate path in a rare chance of congestion issue in this specific case. But when we increased mesh nodes from 9 to 16, then the alternate path options were greater than the previous case. Here, the device maximum data-rate is same as in previous case. However, the relay intermediate nodes can be congested as they have higher degree of connectivity and multiple nodes can transmit data using single relay because of best path selection mechanism. In this case, when we have 100Kbps application data-rate, we observed little improvement because queued data remained below the threshold due to less application rate. This performance is changes when we increased mesh nodes in grid from 25, 36, 49 and 64. Due to increase in this number, there is also increase in data disseminating nodes, therefore, there is also an increase in network throughput. We observed that it increases while node density moves from sparse to dense mode. Both protocols graph slop shows the same trend, but the performance of CA-HWMP is better than the performance of HWMP as CA-HWMP utilizes the option of second best path in case if best path is congested. Considering another scenario to observe network throughput with the increased value of application data-rate, which is increases from 100Kbps to 150Kbps. The increment in number of nodes is same. The graph in Figure 6 (b) shows the throughput for this scenario. The simplest case is with 4 number of nodes, and transmission is simplest because all nodes are in the vicinity of one another. In the same scenario, the second case is with 9 mesh nodes and there is still less chances to use of alternate paths, and device rate is also 3 times greater than the application rate. Therefore throughput observed using both routing protocol is almost same. When mesh nodes are increased to 16, in case of both protocols, the availability of second best path also greater than previous case. As application data-rate is greater than the previous scenario, the relay nodes have still have greater margin of data forwarding in one time. The only possibility of packets drop from queue is, When queue becomes full. But if we continue to increase in number of relay nodes in grid, the graph shows the increasing throughput degradation behavior while using HWMP. The case when we have 36 nodes in a grid, with the increase in mesh nodes in grid, the availability of alternate path also increases. The best path also can have few hops to reach destination, and relay nodes may have to forward data on behalf of few neighbors. At relay node, the multiple flows can result in dropping packets from the queue. This packet drop ratio increases when more nodes enter in network to communicate. This degradation is even more in case of 49 nodes as compared to 36 nodes in grid, When we added more nodes in network, the new entering nodes generate more data to send/share in the network. However, by increasing mesh nodes in the network, there is also a increase in control overhead because of exchange of control messages. Furthermore, there is also an increase in contention of channel access, along the increase in frequency of data collisions and retransmissions. A relay node may drop packets due to queue overflow because of greater traffic load through them with the increase in node density. The graph in Figure 6 (b) presents two graph lines for CA-HWMP and HWMP, where CA-HWMP performance in term of throughput is better than HWMP. The reason is advantages of alternate paths at relay node, when an already selected path gets congested. In case of 64 nodes, this gain is maximum because of available alternate paths are also greater than previous cases which have less node density.

The above case discussed the 150Kbps application rate with the variation of node density. Now considering another graph in Figure 6 (c) with the increase in application data-rate i.e. 200Kbps. The simulation scenario continued with the increment of relay nodes as considered in previous scenario. Th graph lines shows the visible degradation of throughput with the use of HWMP in the whole scenario. First two cases, with 4 and 9 mesh nodes in network, the both protocols i.e. CA-HWMP and HWMP performs similar, But when the relay nodes vary from 16 to 49, the degradation of throughput is visible because of increment in number of hops and relay traffic. In this case, where application has 200Kbps data-rate and a relay node can transmit with maximum 350Kbps data-rate. The relay node has only capacity of transmitting data two nodes with current data-rate, and it queues the remaining data. Ultimately, the queue becomes full and drops packets from it. Results show that with the increase in nodes and intermediate hops, the throughput degraded due to overhead of control messages, collisions, and buffer overflow. Both protocol showed this trend, however, CA-HWMP shows the 13-18% gain in throughput depending upon network traffic.

In the next scenario, we increased data-rate from 200Kbps to 250Kbps, and the graph in Figure 6 (d) show the observed throughput with variation of node density. The first two case with 4 and 9 nodes are same as discussed in above scenario. In this scenario, an application sending data-rate is 250Kbps and node maximum sending data-rate is 350Kbps. In case of 9 nodes, the one node is in the junction of all other nodes, which can be in the best part of each node involving one hop. When it receives data from multiple nodes, initially it queues data, then eventually drops it from the queue due to buffer overflow. It drops to network throughput. In case of CA-HWMP, when number of nodes varies from 16 to 64, its performance is better than HWMP due to alternate paths availability. The traffic is re-routed to alternate route, if the

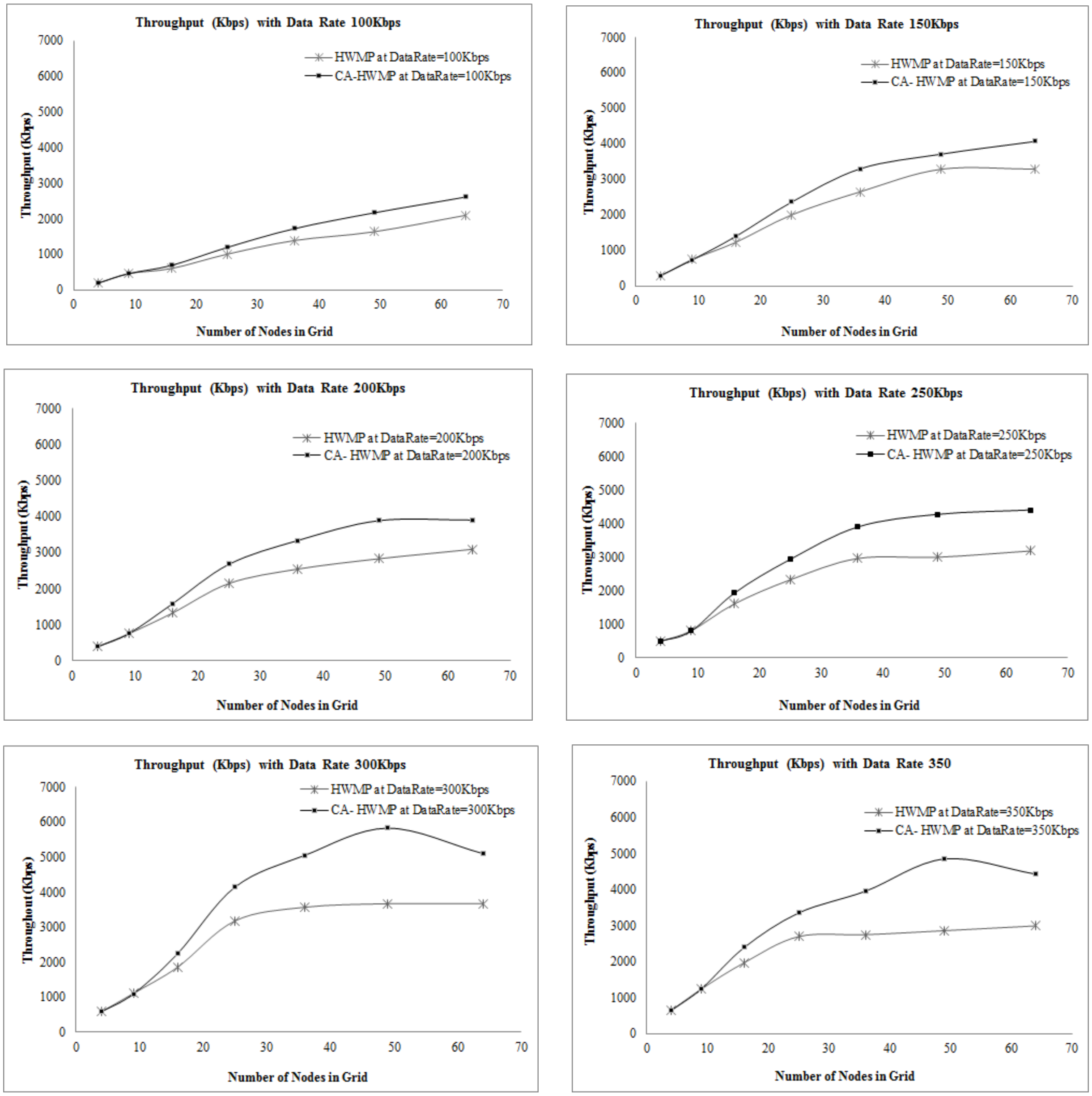


Fig. 6: Throughput Comparison of HWMP and CA-HWMP at (a) 100Kbps; (b) 150Kbps;(c) 200Kbps;(d) 250Kbps; (e) 300Kbps;(f) 3500Kbps

first best path is congested. CA-HWMP graph line shows that when we increase the number of nodes in a grid from 49 to 64, this improvement is almost equal to as we achieves in 49 node grid. Although, CA-HWMP gives possibility of traffic re-route, but when nodes in the network and generated traffic increase upto a level, the high control messages, collision and interference reduce the overall performance of the network. In any case, CA- HWMP is performing better than HWMP.

In Figure 6 (e), the graph shows the trend of throughput in the network with 300Kbps application daterate. In the case of 4 nodes, the performance of both nodes is similar. But with the increase in the mesh nodes from 4 to 9, throughput drops due to network topology and middle relay node selection for mostly best path. In this case, application data-rate is almost equal to the device data rate, therefore relay has the capability to transmit one node data at one time, and if it has to transmit more than one node data, the it maintained queues. This can lead to queue to overflow. As there are less possible paths in case of congestion, hence there is no real advantage of using CA-HWMP in this scenario. Both protocols perform almost likewise. The gain in throughput is visible, when we increase relay node nodes from 9 to 16 in case of CA-HWMP. The gain in throughput is more visible, when further increase nodes i.e. 25, 36, 49. The graphs show that with the increase from 36 to 49, performance difference between both protocols is most significant. Considering HWMP in this scenario, the performance drops due to relay traffic at relay nodes. In the same scenario CA-HWMP performs good, because it shifted traffic on the alternate second best path. CA-HWMP performance is on peak with 49 nodes in network. At the same time, the graph also shows the stable performance with 64, and the gain is less than the 49 node's grid. Although CA-HWMP exploits the possibility of alternate paths, but increase in the mesh nodes also increase control messages, channel contentions and re-transmissions.

In the last scenario, we fixed application data-rate exactly equal to the device data-rate to observe the behavior of network. First two scenario are same, as discussed in previous cases. In the third scenario, when we increase mesh nodes to 16, both routing protocols perform good, but CA-HWMP performance is better than HWMP. When mesh nodes vary from 16 to 49, HWMP performance degraded. But CA-HWMP performs better than HWMP, due to the benefits of alternate paths, and re-route traffic on second optimal path when first gets congested. With 64 nodes in the network, throughput degrades in scenario of HWMP protocol because with the expansion of the network, the number of hops also increase between source and destination (Assuming first node is a source node and last node is the destination node). Therefore, the relay nodes may drop packets from queue as arrival rate is greater than the device forwarding data-rate. Though, CA-HWMP is performed better, but the gain in throughput is not good. Its performance is also affected by exchange of control messages, routing protocol messages and channel contention. The graph in Figure 6 (f), shows that CA-HWMP is performance is better than HWMP, it also shows the decline in throughput when we increase the number of nodes in grid 49 to 64, even in case of CA-HWMP. In case of 49 nodes grid, CA-HWMP performance is most significant.

We observed the network throughput by running simulation of number of scenarios. It is observed that the network throughput is increased with the increase an application data-rate. We also observed that the performance of CA-HWMP is better than HWMP with the increase in application data-rate. But the increase in throughput is limited to specific number of hops and network traffic. We observed maximum throughput when we have 300Kbps application data-rate and 49 mesh nodes in the grid. All graphs show that the performance of CA-HWMP is better when the number of nodes in network are between 25 to 64 as compared to HWMP and the maximum gain is at 300Kbps data-rate in the network of 49 node grid. This shows that the performance is limited to the limited hop count for congestion avoidance.

B. Effect of Application Data-rate on Packet Delivery Fraction(PDF)

The second evaluation paramer, that we consider for evaluation of our proposed mechanism is Packet Delivery Fraction (PDF) to examine the network behavior on increasing data-rate. This PDF is achieved by computing percentage receiving data-rate at receiving nodes. Figure 7 from (a) to (e) show the graphs of PDF for both considered protocols. In the graphs, X-axis represents the number of nodes in the network, while Y-Axis represents percentage value of PDF. By using these values, we monitor the actual gain in proposed technique.

When the number of nodes increases in the grid, the PDF decreases due to increase of intermediate hops, contention for channel access and control overhead. In-spite these factors, with the increase in flows, the PDF decreases on relay nodes due to buffer overflow in absence of congestion control mechanism in HWMP.

Consider a scenario, where application has data-rate of 100Kbps, device transmitting rate is 350Kbps and relay nodes vary from 4-64. The graph in Figure 7 (a) shows that both protocols has decreasing behavior on increasing number of mesh nodes. However, this trend is less in case of CA-HWMP. First two cases are similar to the throughput case. The forwarding rate of device is more than three times of application data rate and mostly nodes are in direct access of each other, therefore this decrease is normal in both protocols. In case of 4×4 Grid, this ratio drops due to increase in hop count between source and destination and also because of other factors like channel contentions, collisions and re-transmissions. With the increase of mesh nodes in the network, this drop also increase. The best path selection mechanism in HWMP is hop count. In this scenario, nodes only select path with minimal hops and ignore the device transmitting capacity and queue length. Mostly traffic pass though the shortest path, in few scenario a single node becomes a bottleneck, which leads to packets drop. In CA-HWMP, nodes consider queue length during path selection and utilize alternate path as well, if congestion occurs during transmission. Results also show that CA-HWMP helps to improve gain in PDF. In the next scenario, the considered topology as used in previous scenarios. We kept all all other parameters constant and only changed application data-rate 150Kbps. The graph lines in Figure 7 (b), presents no significant difference observed between both protocols when we have 4 and 9 nodes grid. With the increase in mesh nodes from 9 to 16 along 50% traffic flows, the CA-HWMP PDF gain is greater than HWMP due to the greater possibility of alternate paths as to $2 * 2 and 3 * 3$ grid. In case of 25 mesh nodes, though more nodes are available to disseminate data, however PDF gain in CA-HWMP is better as compared to HWMP. The simulation results show that 15% packets are dropped due to buffer overflow and 12% packet lost because of wireless reasons HWMP. With the use CA-HWMP this PDF increases 14%. In a scenario with 36, 49 and 64 mesh nodes, decrease in PDF is greater than the previous scenarios. In case of 64 mesh nodes, this PDF degradation is observed 50%. The reason is that packets drop from the queue due to greater difference between incoming and outgoing traffic rate. But the buffer overflow on relay nodes is not only reason, packets also drop due to wireless channel access, interference, packet collisions. However, the gain in PDF is improved in all scenarios as compared to HWMP.

The gain in the PDF shows that performance of mesh network is improved with the use CA-HWMP instead of HWMP. This improvement is achieved by utilizing alternate best paths when the best pat is blocked due to network congestion. But overall decline in PDF with the increase of network diameter is also because of channel contention, control messages overhead, and increase in number of hops. Consider another scenario and graph shown in Figure 7 (c) where the variation in network topology is same but the application data-rate is changes to 200Kbps. All other network parameters are kept constant including application data rate, and we observed the PDF but we changes the grid topology of 4 and 9 nodes, the PDF value in both routing protocols is almost same. In

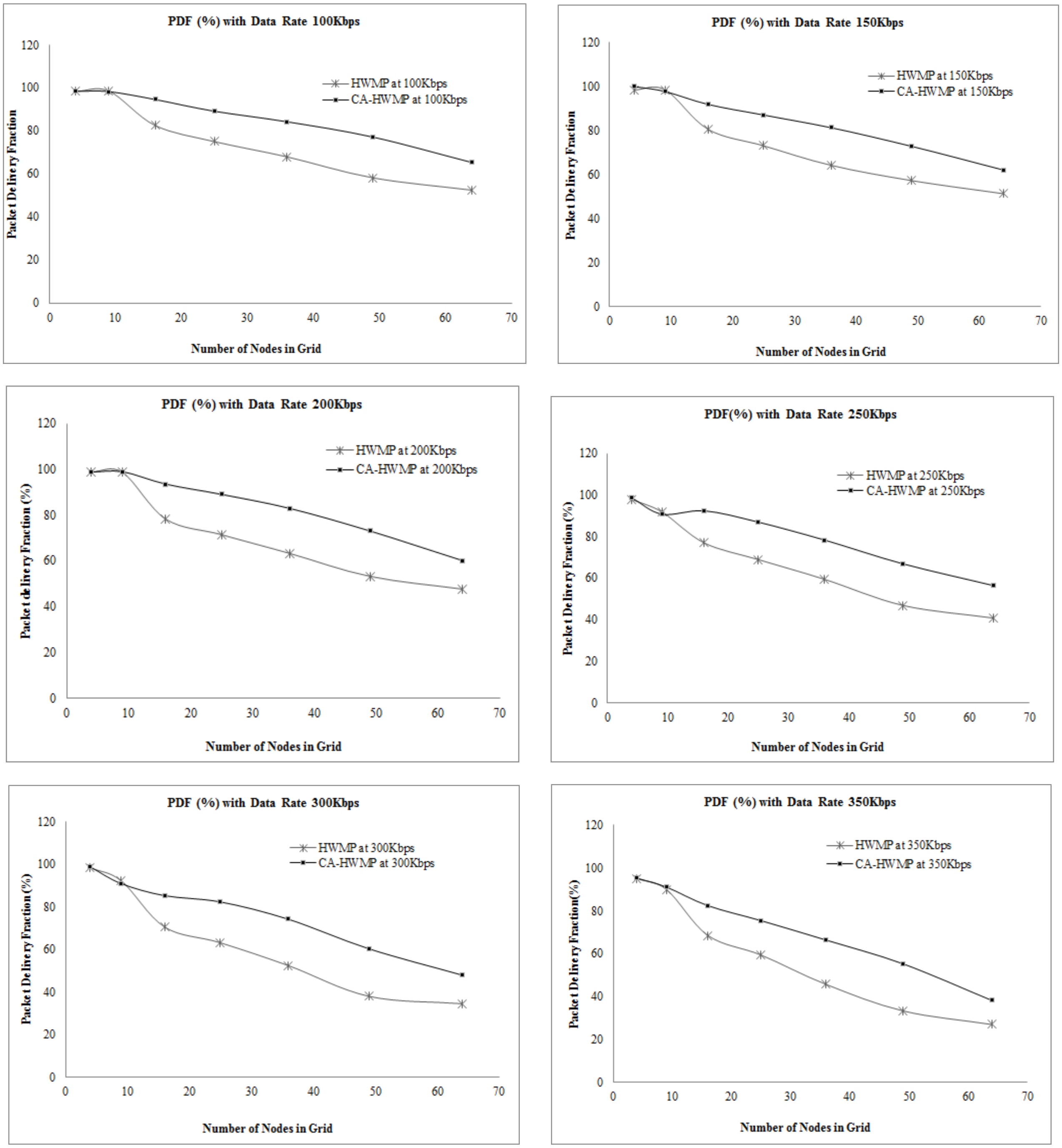


Fig. 7: PDF Comparison of HWMP and CA-HWMP at (a) 100Kbps; (b) 150Kbps;(c) 200Kbps;(d) 250Kbps;(e) 300Kbps;(f) 350Kbps

case of HWMP routing protocol, by increasing mesh nodes to 16, this ratio decreases to 76% due to increase hop count between sender and receiver. When relay nodes receive data from multiple nodes, if this receiving rate is greater than the transmitting rate, they queued received data, later this situation leads to congestion. However, decline in PDF is not because of congestion. With the increase in network nodes, the network traffic load also increases, control messages traffic and due to interference, the chances of collision also increases. All these factors lead together in PDF decline. When we have 16 nodes in the network, the statistics show that there is about 15-17% packet lost due to congestion while remaining packet lost is due to other factors. When relay nodes in the network varies from 25 to 64, this decline increases due to mentioned reasons.

In case of CA-HWMP in the same scenario, we observed significant improvement in PDF value. Although, PDF decreases with an increase in relay nodes, however this decline is less than HWMP. In these scenarios, the PDF gain increases due to possibility of alternate paths. However, graph in Figure 7 (c) shows that CA-HWMP performs better when mesh nodes varies from 16 to 64, and it performs best when there are 49 node network with 50% traffic load.

Consider another scenario where we fixed application data-rate 250Kbps, but mesh nodes vary from 4 to 64 with 50% network traffic. First scenario with 4 nodes grid is always simple. In case of 9 mesh nodes, topology allows one in junction, can be a bottle-neck due to data forwarding. In this scenario, the device transmitting rate is 350Kbps, while application rate is 250Kbps where relay nodes maintain queues to handle incoming data. At one point, these queues get congested and may result in packet loss. When we have 16 mesh nodes in the network, the PDF value decreases to 76%, and which further increases by increasing mesh nodes. However, the simulation results show that CA-HWMP performs better than HWMP due to possibility of alternate paths, this difference is visible in Figure 7 (d).

The graph line of CA-HWMP in Figure 7 (d) shows the increase in PDF gain when new nodes enter into the network, but this gain decreases when we have 64 mesh nodes in this scenario. With the increase in mesh nodes in grid, there is also increase in node interference, channel contention, packet collisions and control overhead, which lead to the decline in PDF.

Consider a scenario with 300Kbps application data-rate and nodes vary from 4 to 64. The performance of both protocols is good in the grid of 4 nodes. The decline in PDF is observer when we increase mesh nodes to 9 and further, where this value decreases due to multi-hop. The queue management issues, channel contention, interference on each relay nodes effects the PDF. The graph lines of CA-HWMP and HWMP in the same scenario show that the performance is similar, because in this scenario there is less possibility of alternate path in case of congestion. When we vary node density by adding new nodes in the network i.e. 16, 25, 36, 49 and 64 nodes, the decline in PDF is greater in both routing protocols. However, this decline in CA-HWMP is less than HWMP. The CA-HWMP performs only better, when there is availability of second best path when already existed becomes congested. Therefore, we observed this PDF gain with increases in mesh nodes.

In the graph shown in Figure 7 (e), X-axis represents nodes the network and Y-axis represents nodes PDF. The graph presents the decline in PDF with the increase in mesh nodes in the network while using both routing protocols. But this decline is greater in HWMP as compared to CA-HWMP. In this scenario, we set application data-rate 300Kbps, and varied grid topology of 4 to 64 nodes. This scenario include application data-rate is almost equal to the device transmission rate i.e. 350Kbps. In the first two scenarios, both protocols perform alike. The simulation results show that when we increase mesh nodes from 9 to 16, 25, 36, 49, 64 in the grid, PDF decreases due to greater data-rate. The relay nodes receive data from multiple nodes, when their receiving rate becomes greater than the device transmission rate. Therefore, the queue becomes full and drops packet. In Figure 7 (f), the graph lines of both protocols show the same behavior that we observed in the last scenario.

Through the simulation results of these multiple scenarios, we concluded that the PDF values depend on an application data-rate and maximum limit relay node to transmit data. If the application generates data-rate greater than the device maximum transmission-rate than packets drop at the application layer. If the application data-rate is less or equal to the device transmission-rate then no packet drop at application layer. In the evaluating scenarios, if there is only one hop involved between source and destination, then there is less probability of packet drops. With the increase in involved relay nodes, this ratio increases due to multiple factors which may include queue, channel contention, interference and control overhead. In our scenario, sending and receiving nodes are chosen randomly, MPs forward data for these nodes. When a MP forwards data on behalf of more than 2-3 nodes then the queue may reach to maximum level then the packets drop from it, resulting in decline of PDF in the network. However, CA-HWMP behaves different from HWMP. It monitors queue level, when it reaches specified threshold, it re-routes data on alternate path. Therefore, there is improvement in PDF while using CA-HWMP. The graphs shown Figure 7 are drawn for multiple scenarios with at different data-rates and they represent PDF difference while using CA-HWMP and HWMP.

C. Effect of Application Data-rate on End-to-End Delay

End-to-end delay is one of the important evaluation parameter. In multi-hop networks, where nodes relay data on behalf of neighbors, they maintained queues. Due to wireless medium, each relay node waits for channel access to transmit data, which adds delay in the packet deliver. To observe the delay in the network, we used different application data-rate with different node density. We computed end-to-ed delay (s) while varying data-rate from 100Kbps to 350Kbps and mesh nodes from 4 to 64. The considered parameters are listed in Table I. We used grid topology with the dimension of $m \times n$ as presented in Figure 5. The ”m” in the grid is number of mesh nodes on X-Axis and ”n” on Y-Axis. In these scenario, we have $n = m = 1, 2, 3, 4, 5, 6, 7, 8$, hence, the grid varies from 1×1 to 8×8.

Consider a scenario with 100Kbps application data-rate and mesh nodes varies from 4 to 64 in the network. In teh first sub-scenario with 4 nodes of grid, and these are directly connect access to one another, therefore observed delay is negligible. When we increase nodes in the grid, the increase in delay is greater than previous scenario due to increase in hop count, channel contention, collisions and re-transmissions as depicted in Figure 8 (a). In the next scenario we increased data-rate to 150Kbps,and the Figure 8 (b) presents end-to-end delay for HWMP and CA-HWMP protocols. The bar lines indicate that the delay increases with the increase in number nodes. Initially, when grid has less nodes in the network, the observed delay was also less, but with the increase in nodes, the observed delay is greater as compared to last scenario. The reason is that, with the increase in nodes in the network, the number of hops for packet traversing are also increases, and if a relay node is responsible for forwarding data on the behalf of multiple node, then due to maintained queue this delay also increased. If the device transmission rate is less than the receiving-rate then more delay added for transmission. The nodes queued packets, and forwarded when they have channel access to transmit it. The queued data increases delay, and if multiple hops are

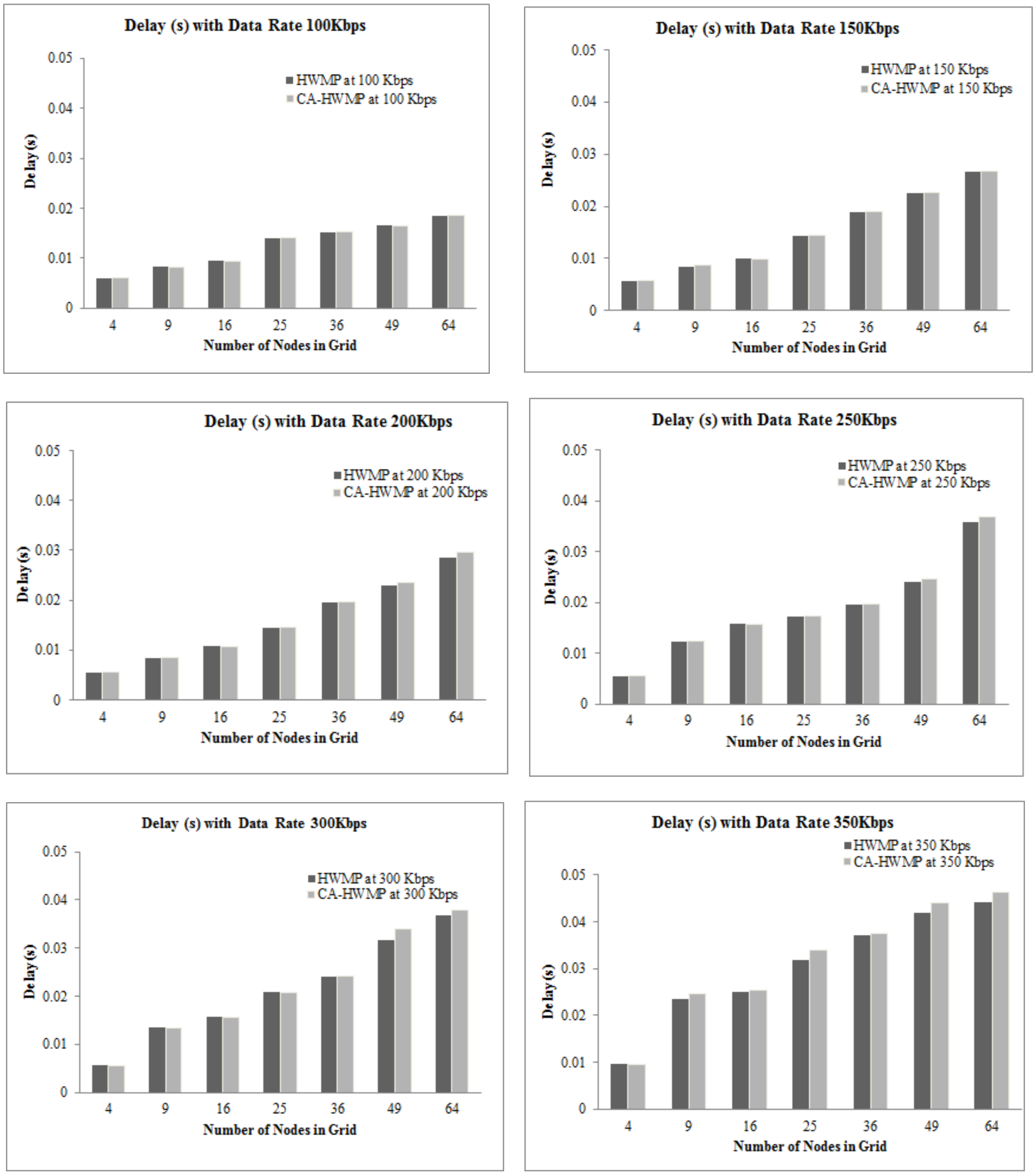


Fig. 8: (Delay Comparison of HWMP and CA-HWMP at (a) 100Kbps; (b) 150Kbps;(c) 200Kbps;(d) 250Kbps;(e) 300Kbps;(f) 350Kbps

involved for transmission then the additional delay is added in the network.

After observing delay when we have application data-rate 150 Kbps, we increase data-rate from 150Kbps to 200Kbps. By increasing the number of nodes in the network, delay increase due to increase in contention, number of hops. Figure 8 (c) shows that observed delay in both protocols is same or negligible greater delay in the case of CA-HWMP. When we have 49, and 64 nodes in the network, packet lost ratio increases due to contention, collisions, re-transmissions and increase in number of hops, therefore, delay observed in HWMP is less. CA-HWMP performs better, and increase packet delivery fraction, as delay calculated on receiving data packets. Therefore greater delay observed in CA-HWMP, however this delay is negligible.

In the next scenario, the application data-rate is 250Kbps and with varied grid topology. In simulation results, we observed that delay is greater than all previous scenarios. With the increase of nodes in the network, the delay also increases because more nodes shares data when new nodes enters in network. These nodes also increase channel contention, number of hops to reach destination where channel contention on each relay node adds delay. Figure 8 (d) presented this discussed trend. As we observed that PDF of CA-HWMP is greater than HWMP, therefore delay is also computed on received packet.

Here, we observed greater delay in case of CA-HWMP and less delay in case of HWMP. In case of CA-HWMP, when queue data stay longer in queue, and its level reached upto specified level, it re-routes data to avoid congestion and increase PDF minimum delay. CA-HWMP chooses second optimum path, which may add delay but this increase in delay is very small.

In this secnario, the application data-rate is 300Kbps and relay nodes vary from 4 to 64. In this scenario like previous simulation scenarios, when new nodes enters in the network, the delay is also increases. In Figure 8 (e), bar lines of both protocols indicate insignificant differences in delay. The difference is visible when we have 49 and 64 mesh nodes in the network in both protocols, in case of CA-HWMP is greater than HWMP. it is computed on the destination node, where PDF is less in HWMP as compared to CA-HWMP. However, difference in delay is much smaller that can be ignored.

The Figure 8 (f) represents the delay chart for both protocols while varying number of nodes in the grid. In the given chart, bar lines of both charts show insignificant difference of the delay between two compared protocols. In case of CA-HWMP, the received packets are greater than HWMP. End-to-end delay computed on the received packet at the destination using the ratio of total delay and total packet received. It is computed by using the time difference of the time when it is sent and the time when it is received at destination. In CA-HWMP, we observed greater values of PDF with ignorant delay.

From Figure 8 (a) to (e), it is concluded that end-to-end delay increases by increasing relay nodes in the the network. With the increase in number of nodes into the network, the throughput of the network also increases due to more disseminate of data into the network, but if the traffic data is more than the network capacity, it introduces congestion in the network. This increase in relay nodes in the mesh grid, also increases interference, packet collisions and channel contention. The increase in the wait time on each relay station also added delay. We fixed all other parameters in both protocol scenarios, and change application data-rate and network density mode. We observed that when with the increase in application data-rate and relay nodes in the grid, the delay also increases due to wireless medium access issues. Packet remained in the queue until node gets a chance to transmit packets.However, the delay difference between both protocols is ignorant.

V. Conclusions and Future Work

The use of Internet with mobility support is at a rapid pace. WMN is one of the wireless technologies which offers high bandwidth and caters mobile users. It has the attractive feature of self-configuring, self-healing and self-organizing and is a suitable candidate for network provisioning in the areas where connectivity through wired media is comparatively difficult or lengthy process. Furthermore, the WMN is a good choice in many scenarios satisfy efficient, where other technologies cannot provide full support.

IEEE 802.11s is a MAC standard with the MAC enhancement in 802.11 MAC for WMN including enhancement of QoS, path selection, security, configuration, and management. It is first IEEE standard, which proposed a routing protocol at the MAC layer i.e. HWMP. It is the mandatory protocol and offers the advantages of both reactive and proactive approaches.

In wireless network, packet collision is generally because of the wireless communication issues, such as contention for channel access, delayed packet in queue due to long wait etc. In IEEE 802.11, one node can only transmit data at a time, due to shared channel characteristics. This restriction adds a significant delay with the increment of the number of hops. The increase in channel contention delay, and queue length leads to congestion. In a WMN, as traffic is aggregated at the MP, MPs near MPP have greater traffic load as compared to other nodes. Therefore, in the absence of any congestion control mechanism, nodes at the outer edges of the network undergo low throughput and increased packet loss. To solve this problem many researchers have proposed algorithms like Total Congestion Control (TCC), Link Selective Congestion Control (LSCC) and Path Selective Congestion Control (PSCC), but every technique have their own pros and cons. In TCC, on receiving CCNF, STA stop sending data to all neighboring nodes although STA can send data to other neighboring nodes. This approach wastes the available bandwidth and add delay. In PSCC, on receiving CCNF the STA only block sending data for the specific link but can receive data. In this scenario, when congestion occurs in the network, STA when it gets congested, it cannot broadcast CCNF message to blocked link. Hence packets drop due to queue overflow. The third is PSCC, which block the specific path on receiving CCNF. For the announcement of specific destination, this algorithm requires modification in the standard CCNF. On receiving modified CCNF a node only block sending data for a specific destination, but it continuously receives data for that specific client. The scenario becomes more complicated when CCNF frame is further broadcasted to immediate node in a continuous chain.

To handle congestion at the MAC layer, we proposed a congestion avoidance technique named Congestion Avoidance Hybrid Wireless Mesh Protocol (CA-HWMP). In this protocol, when node queue level reached to a specified threshold value, it broadcasts CCNF to its immediate neighbors before reaching to congestion state. The nodes present in its neighbor re-route all traffic on congested node from alternate path. For comparison, we have selected our proposed approach using IEEE 802.11s WMN with its mandatory routing protocol i.e. HWMP. For performance evaluation, we used NS3 which is based on object oriented language $C++$ and a scripting language Python. We evaluated our proposed protocol through PDF and average end-to-end delay. We also noticed this effect on the different node grids by gradually varying the environment from sparse to dense mode. From the comparison, it is concluded that CA-HWMP performs better than HWMP in term of greater throughput and PDF. However, CA-HWMP offers negligibly higher delay than HWMP. The increased delay is caused due to selection of alternate path, which may not be the optimal one. However, it offers almost same delay due to congestion as compared to default protocol.

During the evaluation, we found some limitations of our proposed technique and default routing protocol. Scalability is one issue with both protocols. Although CA-HWMP performs better but these both protocols perform good with limited number of mesh nodes. The HWMP did nothing with congestion, but the proposed technique also has limitations which include wait by blocking node in case there is no alternate path to re-route. In such cases, a blocked node waits for the CCNF expiry time to initiate transmission to an already established path which was blocked due to congestion. For the extension of this work, it is recommended to cater the congestion problem in the absence of alternate path. A hybrid technique can also aim to solve this problem. The comparison of our proposed routing protocol with existing congestion control protocol can also consider as its future work.

REFERENCES

[1] Y.-C. Du, Y.-Y. Lee, Y.-Y. Lu, C.-H. Lin, M.-J. Wu, C.-L. Chen, and T. Chen, "Development of a Telecare System Based on Zigbee Mesh Network for Monitoring Blood Pressure of Patients with Hemodialysis in Health Care Centers," *Journal of medical systems*, vol. 35, no. 5, pp. 877–883, 2011.

[2] C. S. Wang and Y.-R. Tzeng, "A Wireless Networking Technologies Overview over Ubiquitous Service Applications," in *Fourth International Conference on Networked Computing and Advanced Information Management (NCM'08)*, vol. 1. IEEE, 2008, pp. 156–161.

[3] V. C. Gungor and F. C. Lambert, "A Survey on Communication Networks for Electric System Automation," *Computer Networks*, vol. 50, no. 7, pp. 877–897, 2006.

[4] Z. Chen, L. Chen, Y. Liu, and Y. Piao, "Application Research of Wireless Mesh Network on Earthquake," in *International Conference on Industrial and Information Systems (IIS'09)*. IEEE, 2009, pp. 19–22.

[5] M. S. Akbar, M. S. Khan, K. A. Khaliq, A. Qayyum, and M. Yousaf, "Evaluation of IEEE 802.11 n for Multimedia Application in VANET," *Procedia Computer Science*, vol. 32, pp. 953–958, 2014.

[6] J. Ishmael, S. Bury, D. Pezaros, and N. Race, "Deploying Rural Community Wireless Mesh Networks," *IEEE Internet Computing*, vol. 12, no. 4, pp. 22–29, 2008.

[7] M. Seyedzadegan, M. Othman, B. M. Ali, and S. Subramaniam, "Wireless Mesh Networks: WMN Overview, WMN Architecture," in *International Conference on Communication Engineering and Networks IPCSIT*, vol. 19, 2011.

[8] M. L. Sichitiu, "Wireless Mesh Networks: Opportunities and Challenges," in *Proceedings of World Wireless Congress*, 2005.

[9] M. S. Akbar, K. A. Khaliq, and A. Qayyum, "Vehicular MAC Protocol Data Unit (V-MPDU): IEEE 802.11 p MAC Protocol Extension to Support Bandwidth Hungry Applications," in *Vehicular Ad-hoc Networks for Smart Cities*. Springer, 2015, pp. 31–39.

[10] T. Sprodowski and J. Pannek, "Stability of distributed MPC in an intersection scenario," in *Journal of Physics: Conference Series*, vol. 659, no. 1. IOP Publishing, 2015, p. 012049.

[11] S. Badombena-Wanta and E. Sheybani, "Mobile Communications for Development: Enabling Strategic and Low-cost e-applications for Rural and Remote Areas," in *IEEE Wireless Telecommunications Symposium (WTS)*, 2010, pp. 1–7.

[12] M. Portmann and A. A. Pirzada, "Wireless Mesh Networks for Public Safety and Crisis Management Applications," *IEEE Internet Computing*, vol. 12, no. 1, pp. 18–25, 2008.

[13] W. Guo and M. Zhou, "An Emerging Technology for Improved Building Automation Control," in *IEEE International Conference on Systems, Man and Cybernetics (SMC 2009)*. IEEE, 2009, pp. 337–342.

[14] A. Yarali, B. Ahsant, and S. Rahman, "Wireless Mesh Networking: A Key Solution for Emergency & Rural Applications," in *IEEE Second International Conference on Advances in Mesh Networks (MESH 2009)*, 2009, pp. 143–149.

[15] K. A. Khaliq, A. Qayyum, and J. Pannek, "Methodology for Development of Logistics Information and Safety System Using Vehicular Adhoc Networks," in *Springer Dynamics in Logistics*, 2017, pp. 185–195.

[16] K. A. Khaliq, A. Qayyum, J. Pannek, "Synergies of Advanced Technologies and Role of VANET in Logistics and Transportation," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 7, no. 11, pp. 359–369, 2016. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2016.071148

[17] X. Wang and A. O. Lim, "IEEE 802.11 s Wireless Mesh Networks: Framework and Challenges," *Elsevier Ad Hoc Networks*, vol. 6, no. 6, pp. 970–984, 2008.

[18] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless Mesh Networks: A Survey," *Elsevier Computer Networks*, vol. 47, no. 4, pp. 445–487, 2005.

[19] A. B. Forouzan, *Data Communications & Networking (sie)*. Tata McGraw-Hill Education, 2007.

[20] R. Karrer, A. Sabharwal, and E. Knightly, "Enabling Large-scale Wireless Broadband: The Case for TAPs," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 1, pp. 27–32, 2004.

[21] "IEEE 802.11s/D8.0," Draft Standard, Tech. Rep., 2009.

[22] H. Aoki, S. Takeda, K. Yagyu, and A. Yamada, "IEEE 802.11 s Wireless LAN Mesh Network Technology," *NTT DoCoMo Technical Journal*, vol. 8, no. 2, pp. 13–21, 2006.

[23] K. Shi, Y. Shu, and J. Feng, "A MAC layer Congestion Control Mechanism in IEEE 802.11 WLANs," in *Fourth International Conference on Communications and Networking in China (ChinaCOM 2009)*. IEEE, 2009, pp. 1–5.

[24] K. A. Khaliq, M. S. Akbar, A. Qayyum, E. Elahi, and A. Zaheer, "Congestion Avoidance Hybrid Wireless Mesh Protocol (CA-HWMP) for IEEE 802.11s," *Elsevier Procedia Computer Science*, vol. 32, pp. 229–236, 2014, The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187705091400619X

[25] K. A. Khaliq, S. Hussain, A. Qayyum, and J. Pannek, "Novel Data Link Layer Encoding Scheme for Multi-hop Wireless Mesh Network," *Procedia Computer Science*, vol. 52, pp. 665–669, 2015.

[26] D. Fu, B. Staehle, R. Pries, and D. Staehle, "On The Potential of IEEE 802.11 s Intra-mesh Congestion Control," in *ACM Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems*, 2010, pp. 299–306.

[27] J. Camp and E. Knightly, "The IEEE 802.11 s Extended Service Set Mesh Networking Standard," *IEEE Communications Magazine*, vol. 46, no. 8, pp. 120–126, 2008.

[28] G. Feng, F. Long, and Y. Zhang, "Hop-by-Hop Congestion Control for Wireless Mesh Networks with Multi-channel MAC," in *IEEE Global Telecommunications Conference (GLOBECOM 2009)*, 2009, pp. 1–5.

[29] A. Raniwala, D. Pradipta, and S. Sharma, "End-to-End Flow Fairness over IEEE 802.11-based Wireless Mesh Networks," in *IEEE 26th IEEE International Conference on Computer Communications (INFOCOM 2007)*, 2007, pp. 2361–2365.

[30] M. Ahmed and K. A. Rahman, "Novel Techniques for Fair Rate Control in Wireless Mesh Networks," *International Journal*, vol. 3, 2012.

[31] S. Rangwala, A. Jindal, K.-Y. Jang, K. Psounis, and R. Govindan, "Understanding Congestion Control in Multi-hop Wireless Mesh Networks," in *ACM Proceedings of the 14th ACM International Conference on Mobile Computing and Networking*, 2008, pp. 291–302.

[32] B. Staehle, M. Bahr, D. Fu, and D. Staehle, "Intra-mesh Congestion Control for IEEE 802.11 s Wireless Mesh Networks," in *21st IEEE International Conference on Computer Communications and Networks (ICCCN 2012)*, 2012, pp. 1–7.

Time Varying Back Propagating Algorithm for MIMO Adaptive Inverse Controller

Ibrahim Mustafa Mehedi
Center of Excellence in Intelligent Engineering Systems (CEIES)
Department of Electrical and Computer Engineering
King Abdulaziz University, Jeddah - 21589, Saudi Arabia

***Abstract*—In the field of automatic control system design, adaptive inverse is a powerful control technique. It identifies the system model and controls automatically without having prior knowledge about the dynamics of plant. In this paper neural network based adaptive inverse controller is proposed to control a MIMO system. Multi layer perception and back propagation are combinedly used in this investigation to design the NN learning algorithm. The developed structure represents the ability to identify and control the MIMO system. Mathematical derivation and simulation results for both plant identification and control are shown in this paper. Further, to prove the superiority of the proposed technique, performances are compared with recursive least square (RLS) method for the same MIMO system. RLS based adaptive inverse scheme is discussed in this paper for plant identification and control. Also the obtained simulated results are compared for both plant parameter estimation and tracking trajectory performance.**

Keywords*—*Adaptive inverse control; neural network; MIMO; multilayer perception

I. Introduction

Prior knowledge is an important factor for almost every conventional control system. Such as in continuous time system, number of poles and zeros or the limit of upper bounds on the order of the plant are assumed to be known [1], [2], [3], [4]. Again the known time delay is crucial for discrete-time systems [5], [6], [7]. To overcome these difficulties, the adaptive control methods were developed. Because it can work even if the system structure and critical parameters are unknown [6], [10]. There are several approaches are proposed to develop the adaptive controllers [8] and already been implemented for different robotic applications. Such an application is presented in [9]. In this example work, an adaptive neural network control approach is used to enhance the performance of a flexible manipulator. Adaptive controllers, based on self-Tuning method were proposed to avoid the problem of uncancellable zeros for the system transfer function [11], but the reference model of the adaptive control depends on transfer function of the plant. Due to this problem, the desired output is not independent of the plant characteristics. The adaptive inverse control is one of the solutions to overcome these difficulties. It is a method to design an automatic control system. It learns over time to control a particular dynamic system [12]. Adaptive filtering technique proceeds with three concurrent learning steps and eventually develops adaptive inverse control method [13]. At the beginning, the modeled adaptive plant identifies the system dynamics. Next, the control dynamics of the plant is learned by a feed-forward controller. Finally, the disturbance affecting the plant is canceled by an adaptive feedback disturbance canceler. These controllers approximately compensate the effect of numerator polynomial at the output with the help of approximate inverse of the plant [4], [10]. The desired trajectory is approximately followed by the output of the plant with some delay which can be estimated.

The plant dynamics is controlled by several neural network (NN) approaches. A dual step controller based on neural-network is used to obtain feedback linearizion and learning of the plant dynamics [14]. The calculation complexities of computing inverse dynamic are reduced by using neural network method. It also improves the precision by learning procedure. A different neural network technique is considered using a feed-forward inverse recurrent method based PD controller [15]. Inversion error is compensated and disturbances are rejected using this technique. Past investigations show the better performance while using neural network controllers for controlling the nonlinear plant dynamics [16]. Gain tuning is also performed for PD controller using NN [17].

All of these techniques used for neural network controllers have firm restrictions. In general, they require to know the fairly precise plant model before hand. Adaptive inverse control technique is considered in this paper, which is based on neural network using multi layer perception for MIMO system. A simplified architecture of the NN models are incorporated in which the modeling of the system approximate inverse of the plant are obtained directly. Then the approximate inverse model is used for the learning process to control the plant dynamics.

The rest of the paper is presented as follows: The problem stated for the purpose of this investigation is mentioned in the the next section. Architecture of adaptive inverse control technique is explained in Section 3 for multi input and multi output (MIMO) system. Multilayer perception based neural network concept is discussed in Section 4. In Section 5, back propagation based learning algorithm is explained to design adaptive inverse controller. Simulation results and their discussion for a dual input and dual output system is presented in Section 6 while using NN based adaptive inverse control scheme. Plant identification algorithm of RLS method and adaptive inverse control scheme is discussed in Section 8. Also the obtained simulated results are compared for both plant parameter estimation and tracking trajectory performance evaluation in this Section. The Section 8 concluded the investigation.

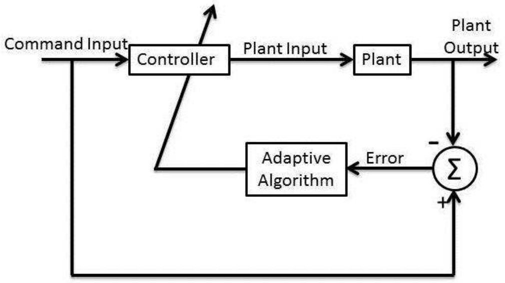


Fig. 1. Schematic diagram of basic adaptive inverse control

II. Statement of Problem

Considering a multi input and multi output (MIMO) discrete time linear system described by:

$$y(z) = P(z)u(z) + V(z) \tag{1}$$

where, multi outputs

$$y(z) = [y_1(z) \quad y_2(z) \quad y_3(z) \quad - \quad - \quad - \quad y_M(z)]^T \tag{2}$$

multi inputs

$$u(z) = [u_1(z) \quad u_2(z) \quad u_3(z) \quad - \quad - \quad - \quad u_N(z)]^T \tag{3}$$

disturbances

$$V(z) = [V_1(z) \quad V_2(z) \quad V_3(z) \quad - \quad - \quad - \quad V_M(z)]^T \tag{4}$$

and the discrete transfer function

$$P(z) = \begin{bmatrix} P_{11}(z) & P_{12}(z) & P_{13}(z) & - & - & P_{1N}(z) \\ -- & -- & -- & - & - & -- \\ -- & -- & -- & - & - & -- \\ -- & -- & -- & - & - & -- \\ P_{M1}(z) & P_{M2}(z) & P_{M3}(z) & - & - & P_{MN}(z) \end{bmatrix} \tag{5}$$

In the above equations, $u(z)$ is the inputs and $y(z)$ is the outputs of the measurable plant while $V(z)$ denotes for bounded disturbances. $P(z)$ is the discrete transfer function metrics. The aim of the neural network based inverse adaptive is to obtain a set of control inputs which are bounded. With the impact of these bounded control inputs, the outputs $y(z)$ should follow the reference inputs.

III. Architecture of Adaptive Inverse Controller

Adaptive inverse controller is not similar to the traditional closed loop controllers. The main concept of inverse adaptive control is to govern the system with a control command from the controller. The controller transfer function is the inverse of plant transfer function. The principal idea of inverse adaptive control is shown in Figure (1). Obtaining better tracking performance for the plant output is the main objective of this system. A true plant inverse need to be created by adapting the controller parameters because the plant is usually unknown. Comparing the plant output and command input, an error signal is produced to use for the adjusting process of the controller's parameters through an adaptive algorithm. Purpose of this algorithm is to minimize the error in terms of square mean. But this configuration has some demerits. Such as the adaptation process of the controller can not be done directly by the algorithm. Because the algorithm (for example, LMS) needs an error refereed to the plant input. Therefore a different configuration of adaptive inverse controller is proposed to overcome this problem and shown in Figure (2) and (3) for a MIMO system.

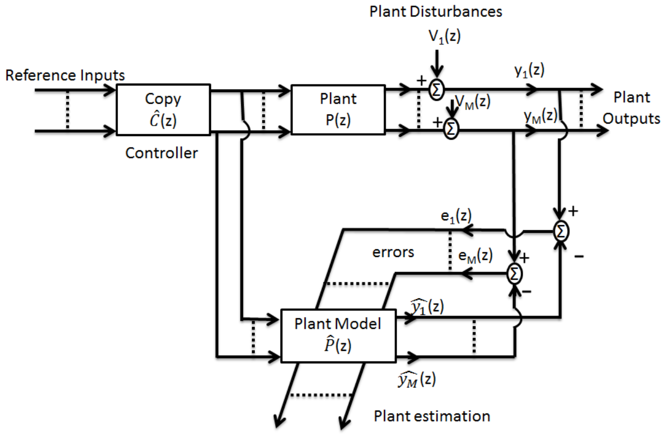


Fig. 2. Adaptive inverse- plant estimation for MIMO system

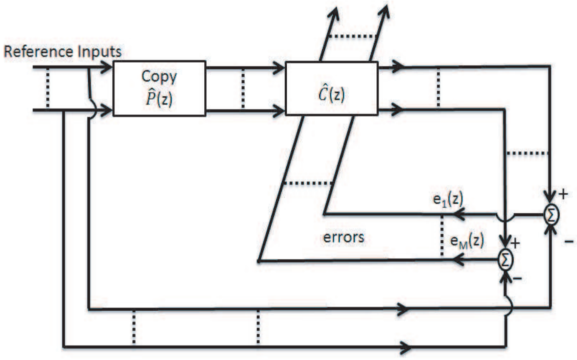


Fig. 3. Adaptive inverse- plant control for MIMO system

Rapid adaptation process and control action with plant disturbance represented in Figure (2) and (3). The plant identification and control mechanism are described as follows:

- Step 1: A MIMO plant model $\hat{P}(z)$ of the original plant $P(z)$ is identified on real time basis by using back propagating adaptive algorithm shown in Figure (2).
- Step 2: Updated parameters of controller $\hat{C}(z)$ is generated from a digital copy of $\hat{P}(z)$ either on-line or off-line and it is shown in Figure (3).
- Step 2: Finally the obtained updated $\hat{C}(z)$ can be used as a cascaded controller with the original plant $P(z)$ as presented in Figure (2).

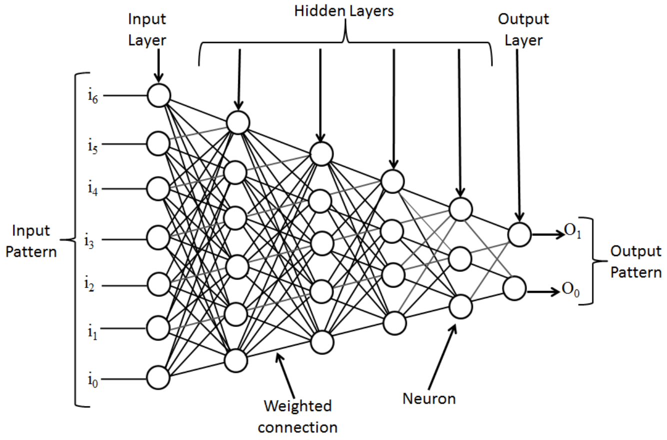


Fig. 4. Multi-layer neural network

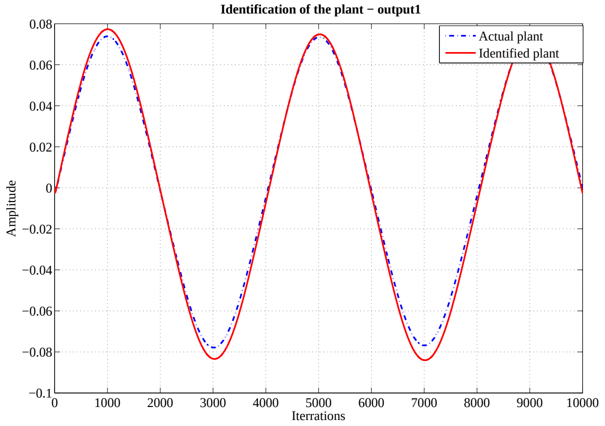


Fig. 5. MIMO plant identification without disturbance for output-1 (Sinusoidal input)

IV. Neural Network with Multi Layer Perception

A multi layer neural network contains multiple neurons, those are organized into different layers . The primary layers are placed at input side, the output layers are organized at the end and the middle area of the input and output is known as hidden layers [18]. It is already known that neural network is capable to execute in the presence of system nonlinearities because NN is a nonlinear filter. This property encourages to implement NN in the adaptive inverse problem. The neurons are connected towards forward direction without having any feed back connections between input and output. Therefore, in this work, the adaptive inverse control is implemented by using multi layer feed forward neural network (MLFFNN) [19] and shown in Figure (4). The active functions of successive layers can differ from each other. Connection link between input and neurons contain some weight. Neuron output is applicable to the nonlinear function.

V. Adapting Controller Via Learning Algorithm

The learning algorithm of this study is supported by back propagation technique for the NN based controller. The activation function induced by local field at the input is shown in Eq.(6). where, $y_i(n)$ is the output of ith neuron for the nth iteration. The synaptic weight of the connecting link between output of ith neuron and jth neuron is denoted by $w_{ji}(n)$. The total number of inputs applied to the neuron j is indicated by m.

$$v_j(n) = \sum_{i=0}^{n} w_{ji}(n) y_i(n) \tag{6}$$

The output at the jth neuron is shown in Eq.(7) while the nonlinear function $\phi()$ is applied to the output of any neuron.

$$y_j(n) = \phi_j(v_j(n)) \tag{7}$$

The synaptic weights is updated according to the back propagation algorithm and it is expresses by the Eq.(8).

$$w_{ji}(n+1) = w_{ji}(n) + \eta(n)\delta_j(n)y_i(n) \tag{8}$$

where δ_j is responsible for local gradient related with jth neuron while the learning rate is denoted by $\eta(n)$. This learning rate is updated using following technique:

$$\eta(n) = \frac{\psi(n) + \psi(n-1) + - - - - - - - - \psi(n-m)}{m+1} \tag{9}$$

here

$\psi = \alpha\eta(n-1) + \gamma \parallel e(n) \parallel^2$.

Mathematical expression of the local gradient δ_j is defined by Eq.(10) for j output neuron.

$$\delta_j(n) = e_j(n)\Phi_j^{'}(v_j(n)) \tag{10}$$

where error e_j is measured between the output and desired response $d_j(n)$. Again the local gradient can be calculated while the neuron comes from hidden layer and expressed by the Eq.(11).

$$\delta_j(n) = \Phi_j^{'}(v_j(n)) \sum_k \delta_k(n) w_{kj}(n) \tag{11}$$

VI. Simulation Results and Discussion

Transfer function of a MIMO system is used for this investigation. To obtain direct and inverse model of MIMO system, we have used back propagating algorithm based on feed forward multi layer perception. The transfer function of double inputs and double outputs MIMO plant is shown in Eq. (12).

$$P(z) = \begin{bmatrix} \frac{z^{-1}(0.1182-0.1531z^{-1})}{1-1.385z^{-1}+0.4724z^{-2}} & \frac{z^{-1}(0.1378-0.1378z^{-1})}{1-1.385z^{-1}+0.4724z^{-2}} \\ \frac{z^{-1}(-0.1174-0.09145z^{-1})}{1-1.385z^{-1}+0.4724z^{-2}} & \frac{z^{-1}(0.09867-0.1683z^{-1})}{1-1.385z^{-1}+0.4724z^{-2}} \end{bmatrix} \tag{12}$$

A. Plant identification

Primarily the system is identified through adaptive inverse back propagation technique with random weight values while no disturbance is considered. Sinusoidal signal is given as the

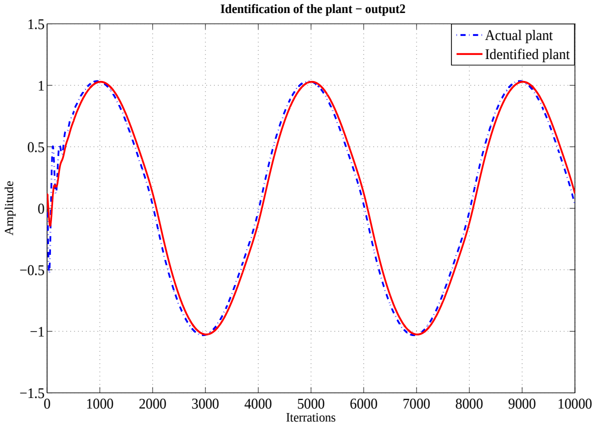


Fig. 6. MIMO plant identification without disturbance for output-2 (Sinusoidal input)

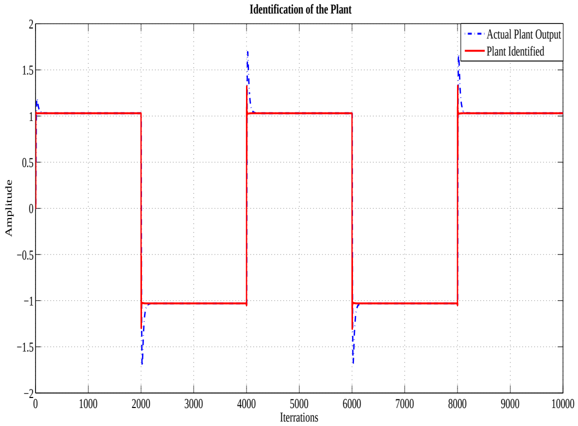


Fig. 7. MIMO plant identification for square input signal without disturbances (sample result)

reference input signal. Identified plant is shown in Figure (5) and (6) with respect to output-1 and output-2. It is observed that the plants are identified perfectly. To see the impact of changing the input signal the simulation was run again using square wave (reference input) as shown in Figure (7). Due to the changes of input signal the proposed technique found the plant identification nearly perfect.

The same simulation was repeated with random weight values in the presence of disturbance. With the same sinusoidal input signal, the identified plant is shown in Figure (8) and (9). Again the sample result of MIMO plant identification is shown in Figure (10) with disturbance condition while the input signal is changed from sinusoidal to square wave.

In both cases, an adaptive inverse with back propagation technique was found the satisfactorily identified system. The system was driven by uniform control signal. It is shown in the Figure (5) to (10) that the neural networks could be trained to identify the plant nearly perfect manner with and without disturbances. Usually with the disturbance, the plant dynamics should be disturbed. With the implementation of adaptive inversed based back propagation technique, the

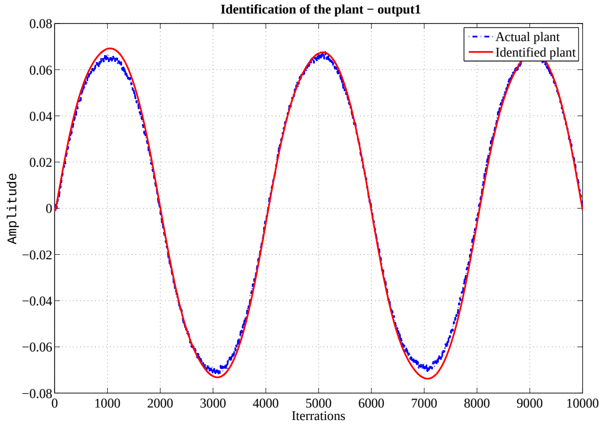


Fig. 8. MIMO plant identification with disturbance for output-1 (Sinusoidal input)

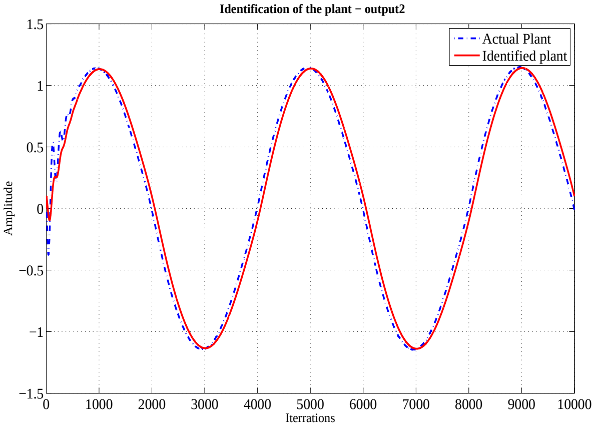


Fig. 9. MIMO plant identification with disturbance for output-2 (Sinusoidal input)

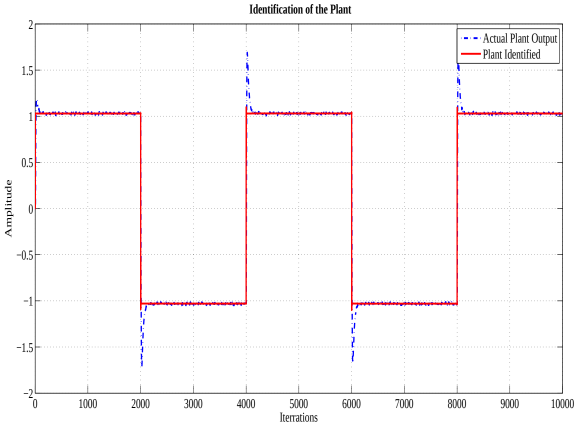


Fig. 10. MIMO plant identification for square input signal with disturbances (sample result)

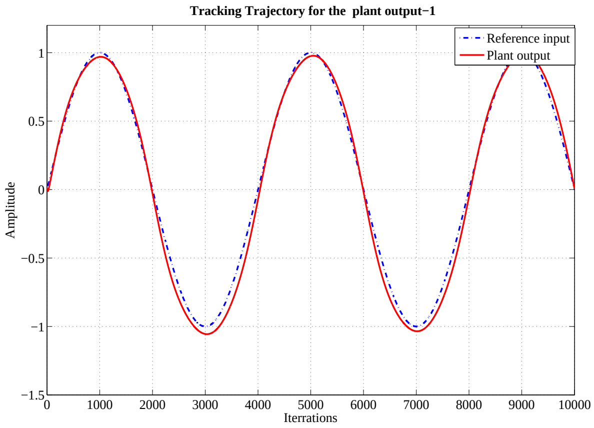


Fig. 11. Tracking trajectory for MIMO plant without disturbance for output-1

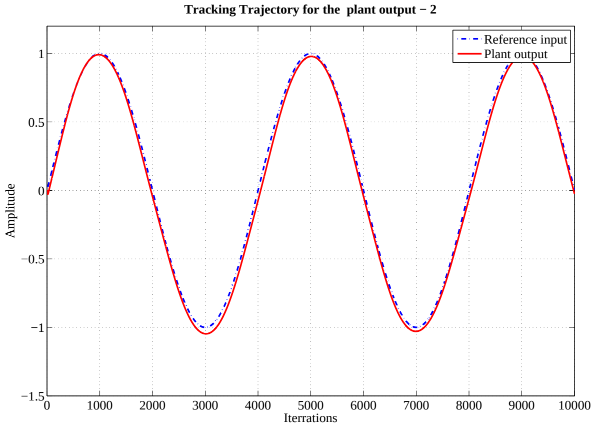


Fig. 12. Tracking trajectory for MIMO plant without disturbance for output-2

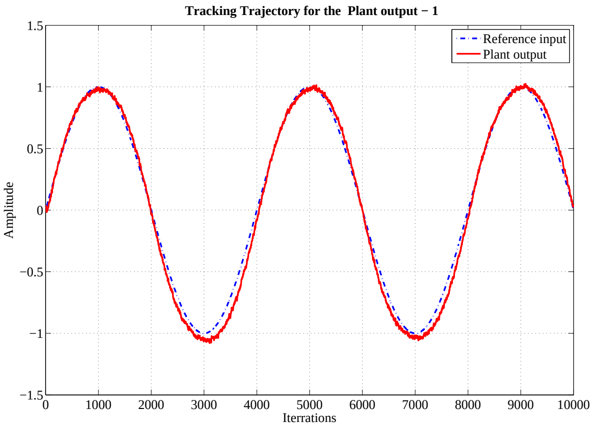


Fig. 13. Tracking trajectory for MIMO plant with disturbance for output-1

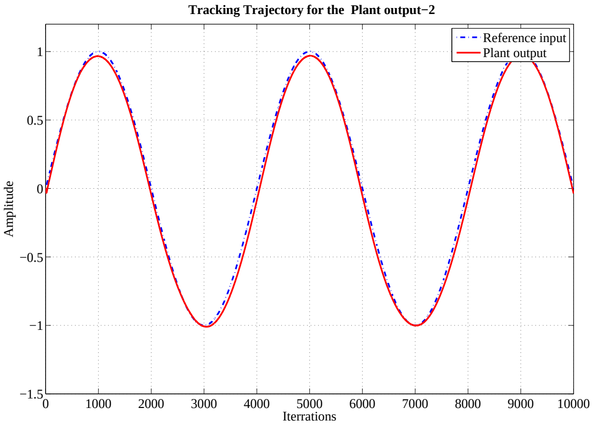


Fig. 14. Tracking trajectory for MIMO plant with disturbance for output-2

system identification processes matched the nominal dynamics of the plant. These proves the theoretical prediction.

B. Plant control

Once the plant identification is done then the control action is implemented using adaptive inverse technique to the MIMO system. Reference input is chosen as sinusoidal signal. Primarily the plant is experienced with no disturbances. The result is presented in Figure (11) and (11). The desired plant output (blue dashed line) and the true system output (red solid line) are indicated in this result. Tracking of the sinusoidal input signal is nearly perfect while the plant does not experience any disturbances. A sinusoidal control signal is observed for this MIMO plant.

To observe the performance of disturbance cancellation, the disturbance signal is included in terms of noise in the algorithm. The filter is chosen for the purpose of disturbance cancellation. The effectiveness of the canceler was tested perfectly. The result is in shown Figure (13) and (14). The control signal of the MIMO plant is also sinusoidal while the disturbance is considered. To see the impact of changing the input signal to control the plant output, the simulation was run again using square wave (reference input) as shown in Figure (15).

VII. Comparison with RLS based Adaptive Inverse Control Algorithm

A. Recursive Least Squares (RLS) Method

The principle task of Recursive Least Squares (RLS) method is to calculate the state variables and observation vectors of the system. Then it compares between observation and the actual output of system. Finally it calculates the sum of squared errors. The parameter matrix is identified through a continuous modification process while the sum of squared error is achieved at its minimum range. Therefore, the identified parameters are kept closer to the actual parameters of the system [20]. Although RLS method is very fast process but it is highly complex in terms of computational cost.

B. Summary of Identification Algorithm

Multi order filter is considered to summarize RLS algorithm. In the Fig. 16, $r(m)$, $y(m)$, $d(m)$ and $z(m)$ are input, output, disturbance noise and measured output respectively.

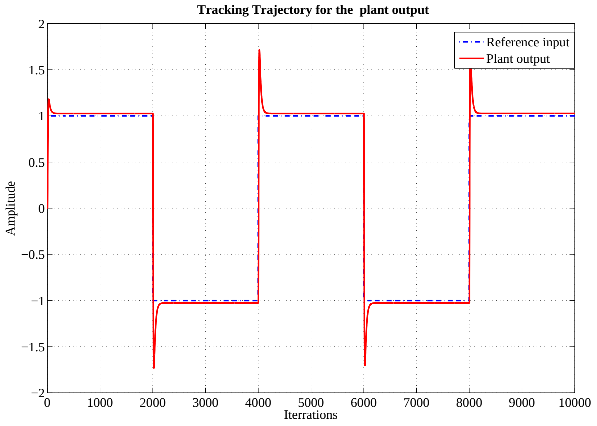


Fig. 15. Tracking trajectory of MIMO plant for square input signal (sample result)

$\alpha(m)$ is model parameter which is unknown. The model input is defined as

$$r(m) = [r_1(m), r_2(m), ...r_n(m)]^T \tag{13}$$

$$\alpha = [\alpha, \alpha, ...\alpha]^T \tag{14}$$

Output parameters of the model is

$$z(m) = r^T(m)\alpha + d(m) \tag{15}$$

The function of least square criterion is deduced by

$$C(\alpha) = \sum_{m=1}^{n} [z(m) - r^T(m)\alpha]^2 \tag{16}$$

α is estimated for the minimum value of $C(\alpha)$ and then $\hat{\alpha}$ is called the parameter values of least square estimation. Now the recursive least square (RLS) method is expressed through

$$\begin{cases} \hat{\alpha}(m) = [r(m)^T r(m)]^{-1} r(m)^T z(m) \\ q(m)^{-1} = r(m)^T r(m) \end{cases} \tag{17}$$

Where $q(m)$ is symmetric matrix positively decrease with the increase of y. The derived formulas for recursive methods are as follows:

$$\hat{\alpha}(m) = \hat{\alpha}(m-1) + M(m)[r(m) - \theta^T \hat{\alpha}(m-1)] \tag{18}$$

Here, $M(m)$ is gain matrix and defined as:

$$M(m) = \frac{q(m-1)\theta(m)}{1 + \theta^T(m)q(m-1)\theta(m)} \tag{19}$$

$$q(m) = [I - M(m)\theta^T]q(m-1) \tag{20}$$

$$C(m) = C(m-1) + \frac{[z(m-1) - r^T(m-1)\alpha(m-1)]^2}{1 + \theta^T(m-1)q(m-1)\theta(m-1)} \tag{21}$$

Therefore, the residual is expressed as:

$$\gamma(m) = \frac{[z(m-1) - r^T(m-1)\alpha(m-1)]^2}{1 + \theta^T(m-1)q(m-1)\theta(m-1)} \tag{22}$$

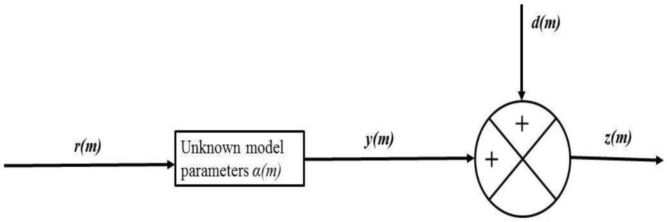


Fig. 16. Least Squares Method

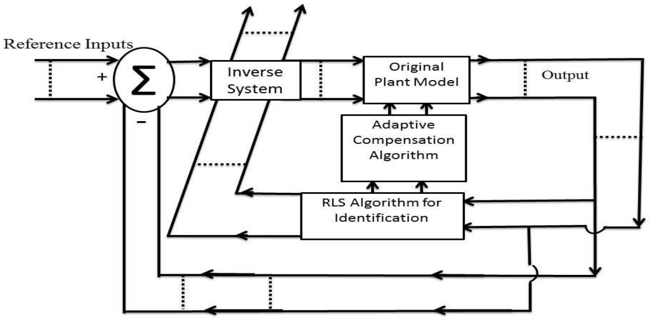

Reference Inputs
+
−
Σ
Inverse System
Original Plant Model
Output
Adaptive Compensation Algorithm
RLS Algorithm for Identification

Fig. 17. RLS based adaptive inverse- plant control for MIMO system

Using suitable initial values for q and $\hat{\alpha}$ recursive operation is performed so that the residual error $\gamma(m)$ is reduced enough. Hence, a minimum value is obtained for criterion function in order to complete the identification process.

C. RLS Based Adaptive Inverse Control

Figure 17 shows a schematic diagram of RLS based adaptive control for MIMO system. Structure contains several blocks like original model of the plant, inverse plant, adaptive algorithm, plant estimation algorithm, and feedback module. In this control architecture, plant is identified using RLS algorithm and expressed into S-function and converted into inverse system which combined with original plant connected in series. State feedback block forms a closed loop control architecture.

D. Simulation results of RLS Based Adaptive Inverse Control - A comparison

AThe same MIMO system defined in Eq. 12 is identified through RLS based adaptive inverse technique with random weight values. A square signal is given as the reference input signal. Identified plant is shown in Figure (18) and (19) with respect to output-1 and output-2. It is observed and compared with the result produced through back propagation based adaptive inverse control shown in Fig. 7 (without disturbance) and Fig. 10 (with disturbance). A better identification for plant parameters is obtained than RLS based estimation technique in terms of overshoot. Specifically, RLS based adaptive inverse algorithm of MIMO plant identification for square input signal with respect to output-2 contains very high overshoot. Therefore, the identified plant using neural network based controller is more perfect over RLS based estimation technique.

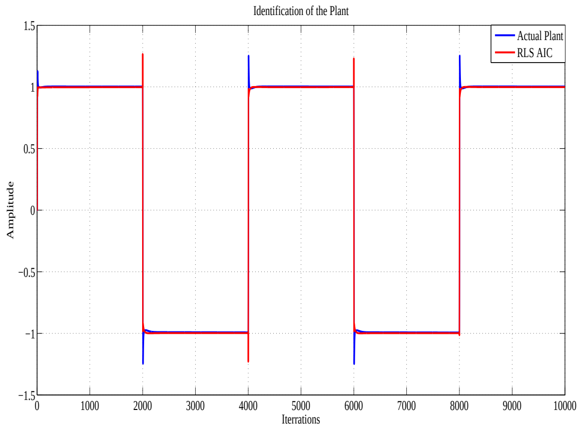


Fig. 18. MIMO plant identification for square input signal with respect to output-1 using RLS based adaptive inverse algorithm

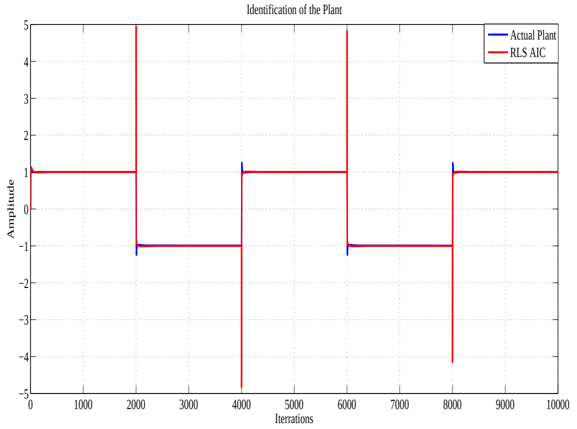


Fig. 19. MIMO plant identification for square input signal with respect to output-2 using RLS based adaptive inverse algorithm

Next step of plant parameter identification is to control the system. RLS based adaptive inverse technique is used to control the outputs of the same MIMO system. Reference input is chosen as square signal for this simulation. The plant controlled results are presented in Figure (20) and (21) with respect to output-1 and output-2. The desired plant output (blue dashed line) and the true system output (red solid line) are indicated in these results. It is observed that using RLS based adaptive inverse algorithm, tracking trajectory of MIMO plant for square input signal with respect to output-2 contains very high overshoot. For the comparison with NN based adaptive inverse control technique, the plant is experienced with no disturbances which is shown in Fig. 15. It is found that the tracking trajectory of MIMO plant is more perfect while using back propagating algorithm based adaptive inverse control technique because its overshoot is with acceptable range. Due to the higher overshoot obtained in Fig. 21, it may cause instability of the system.

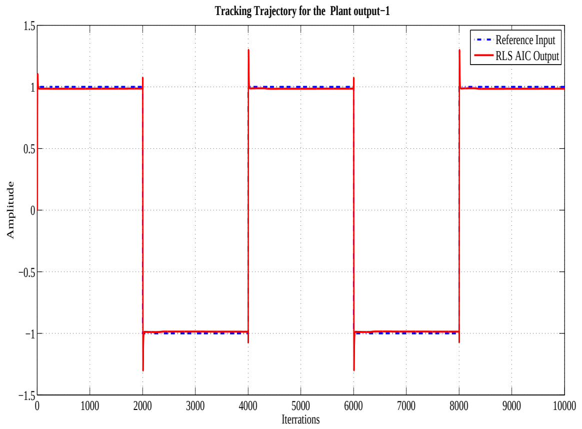


Fig. 20. Tracking trajectory of MIMO plant for square input signal with respect to output-1 using RLS based adaptive inverse algorithm

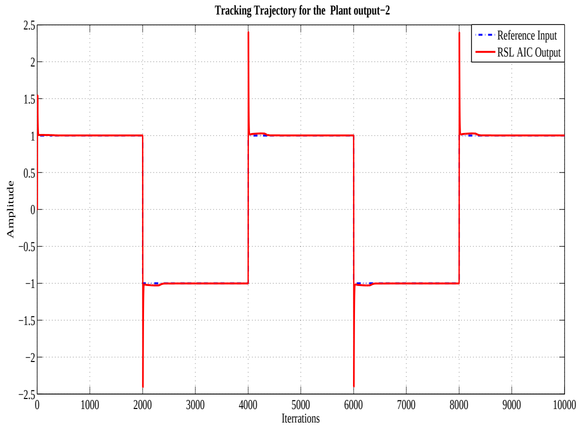


Fig. 21. Tracking trajectory of MIMO plant for square input signal with respect to output-2 using RLS based adaptive inverse algorithm

VIII. Conclusion

In this paper, back propagation based adaptive inverse control technique is proposed to find the approximate inverse of the system. It has been shown that the proposed control method can perform well for MIMO system. It has also been shown nearly perfect performance while the disturbance is injected in term of noise. Therefore, the results verify the ability of neural network based adaptive inverse technique to control MIMO system. To prove the superiority of the proposed technique, the performance is compared with recursive least square (RLS) method for the same MIMO system. Plant identification algorithm of RLS method and adaptive inverse control scheme is discussed in this paper. Also the obtained simulated results are compared for both plant parameter estimation and tracking trajectory performance.

Acknowledgment

This article was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. Therefore, the authors acknowledge with thanks DSR financial support.

REFERENCES

[1] S. Alkhalaf, *Improvement of Control System Performance by Modification of Time Delay*, (IJACSA) International Journal of Advanced Computer Science and Applications, **6**(2), pp. 181-185, 2015.

[2] B. Audone, M. Audone and I. Marziali, *On the use of the minimum phase algorithm in EMC data processing*, International Symposium on Electromagnetic Compatibility (EMC EUROPE), pp. 1-6, 2012.

[3] J. Lu, M. Shafiq and T. Yahagi, *A design method of model reference adaptive control for SISO non-minimum phase continuous-time systems using approximate inverse systems*, Transaction IEEE of Japan, **117-C**(3), pp. 315-321, 1997.

[4] J. Lu, M. Shafiq and T. Yahagi, *A design method of model reference adaptive control for SISO non-minimum phase continuous-time systems based on pole-zero placement*, IEICE Transaction Fundamentals, **E80-A**(6), pp. 1109-1115, 1997.

[5] B. H'mida and S. Dhaou, *Discrete-Time Approximation for Nonlinear Continuous Systems with Time Delays*, (IJACSA) International Journal of Advanced Computer Science and Applications, **7**(5), pp. 431-437, 2016.

[6] K. Astrom and M. B Witten, *Adaptive control*, Addison-Wesely, New York, 1995.

[7] J. Lu, M. Shafiq and T. Yahagi, *A new method for self tuning control of non-minimum phase discrete-time systems in the presence of disturbances*, Transaction IEEE of Japan, **117-C**(2), pp. 110-116, 1997.

[8] A. Peiman, K. Abdollah and H. Khayrollah, *A novel adaptive fully-differential GM-C filter, tuneable with a CMOS fuzzy logic controller for automatic channel equalization after digital transmissions*, AEU - International Journal of Electronics and Communications, **63**(5), pp. 374-386, 2017.

[9] A. R Maouche and H. Meddahi, *A Fast Adaptive Artificial Neural Network Controller for Flexible Link Manipulators*, (IJACSA) International Journal of Advanced Computer Science and Applications, **7**(1), pp. 298-308, 2016.

[10] T. Yahagi and J. Lu, *On self-tuning control of non minimum phase discrete time systems using approximate inverse systems*, Journal of Dynamic Systems, Measurement, and Control, Transaction AMSE, **115**, pp. 12-18, 1993.

[11] K. Astrom and M. B Witten, *self-tuning controller based on pole-zero placement*, IEEE Proceedings. **120-D**, pp. 120-130, 1980.

[12] B. Widrow and G. L Plett, *Adaptive inverse control based on linear and nonlinear adaptive filtering*, Proceedings of International Workshop on Neural Networks for Identification, Control, Robotics and signal/image processing, *Venice, Italy*, pp. 30-38, 1996.

[13] G. L Plett, *Adaptive inverse control of plants with disturbance*, PhD Thesis, Stanford University, Stanford, CA, 1998.

[14] B. S Kim and A. J Calise, *Nonlinear flight control using neural networks*, Journal of Guidance Control and Dynamics, **20**(1), pp. 26-33, 1997.

[15] L. Yan and C. J Li, *Robot learning control based of recurrent neural network inverse model*, Journal of Robotic Systems, **14**(3), pp. 199-212, 1997.

[16] K. S Narendra and K. Parthasarathy, *Identification and control of dynamical system using neural networks*, IEE transaction, Neural Network, **1**(1), pp. 4-27, 1990.

[17] D. L Tien, H. J Kang, Y. S Suh and Y. S Ro, *An online self-gain tuning method using neural networks for nonlinear PD computed torque controller of a 2-dof parallel manipulator*, *Neurocomputing*, **116**, pp. 53-61, 2013.

[18] F. Laurene, *"Fundamentals Of Neural Networks Architectures, Algorithms and Applications*, Prentice Hall, ISBN:0133341860, 9780133341867, 1994.

[19] H. S Adheed and A. Sulaiman, *Multi-Layer Feed Forward Neural Network Application In Adaptive Beamforming Of Smart Antenna System*, International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), DOI: 10.1109/AIC-MITCSA.2016.7759925, pp. 1-6, 2016.

[20] Z. Jianhui and C. Siqin *Adaptive Inverse System Control of Electromagnetic Linear Actuator*, International Journal of Control and Automation, http://dx.doi.org/10.14257/ijca.2015.8.12.12, **8**(12), pp. 131-144, 2015.

Study of the Performance of Multi-hop Routing Protocols in Wireless Sensor Networks

Nouredine Seddiki
Dep of Science, University of Bechar
Bechar, Algeria

Bassou Abedsalem
Dep of Science, University of Bechar
Bechar, Algeria

***Abstract*—Currently in the literature, there are quite a number of multi-hop routing algorithms, some of which are subject to normalization.**

Routing protocols based on clustering provide an efficient method for extending the lifetime of a wireless sensor network. Except that much of the research focuses less on communication between the Cluster-Head (CH), the nodes and the base station, and gives even less importance to the influence of the type of communication on the Life of the network.

The aim of this article is to make a comparative study between some routing algorithms. Since they are not based on analytical models, the exact evaluation of some aspects of these protocols is very difficult. This is the reason why we make simulations.

To study their performance. Our simulation is done under NS2 (Network Simulator 2). It allowed to obtain a classification of the different routing algorithms studied according to one of the metrics such as the loss of message, and the lifetime of a network.

***Keywords*—network sensors; routing protocol; simulation; NS2; d network lifetime**

I. Introduction

For many and various reasons,nowadays and in our daily lives, the evolution of technology ensures that any change in appearance in any environment,is detected, measured and collected through small electronic components called nodes sensors.The sensor nodes have the distinction of being inexpensive compared to traditional sensors and possess limited energy resources for processing, computation and transmission.

In order to study a given phenome No in a given environment, it is necessary to install a wireless sensor network, scattering the ma number of nodes sensors tos can the entire space in question.

In wireless sensor networks, the energy used for the communication of the data captured by nodes,is very high, compared to that used for any other operation.That is why we felt it necessary to give a look to this component.

The minimization of the energy consumption and the extension of the life time of the network, still one of the biggest concerns of researchers in this field.Whya hierarchicaltopology?Whyclustering? And why is a multi-hopcommunication?These are three questions to be answered,in order to underst and completely the choice of protocols sample we collected.

In a flat topology,each node maybe both a sensor, sink and gate way, all nodes communicate with each other, so that communications traffic is very dense and there fore,there is a high consumption of energy, which reduces the lifetime of the network, even up to its exhaustion

To this end, the routing protocols based on groupings of clustered nodes, have been, introduced to provide an effective method to extend the lifetime of a wireless sensor network, and by reducing the communication traffic, which was much denser in a flat topology. First, let us focus only on the network topology. Figures Fig. Ia and Ib Fig illustrate a remarkable reduction in communication traffic between the two topologies, flat and hierarchical[8].

Hence, the choice of the hierarchical topology,where the role of a cluster member nodes can be summarized in the detection information from their environment and their communication to the cluster-head.The cluster-head is for its part aggregates this data and sends it to the base station.

Following the same strategy to achieve the same goal of energy consumption reduction for the extension of the life time of the network, attention,was given to the type of communication used during data exchange between the sensor nodes, Cluster-Heads and the base station, for this their has been a shift from communications ata Single Hop to multi-hop communications. Figures, Fig.I.c, Fig.I.d.and Fig.I.e,present the various possible multi-hop communications in a hie rarchical topology.

Hence, the choice of multi-hop communications.

This document provides a comparative study of some multi-hop routing protocols in wireless sensor networks. It is, organized into four parts as follows:

The first part is a general introduction, in which the characteristics of wireless sensor networks (WSN), are mentioned. The second part is a state of the art: a literature review. As for the third part, it includes a comparison between some protocols and our synthesis drawn from this comparison. To close this document, part fourth, representing a conclusion that justifies that, the chosen protocol, is the most performance in the lot of protocols that we have chosen in our comparative study.

II. Related Works

For the reasons explained above, in this section we will outline some multi-hop routing protocols.

A. KOCA (K-hop Overlapping Clustering Algorithm)

Mustafa A.YOUSSEF Adel Youssefand MohammedF.YOUNIS [2] propose in 2009 an original algorithm

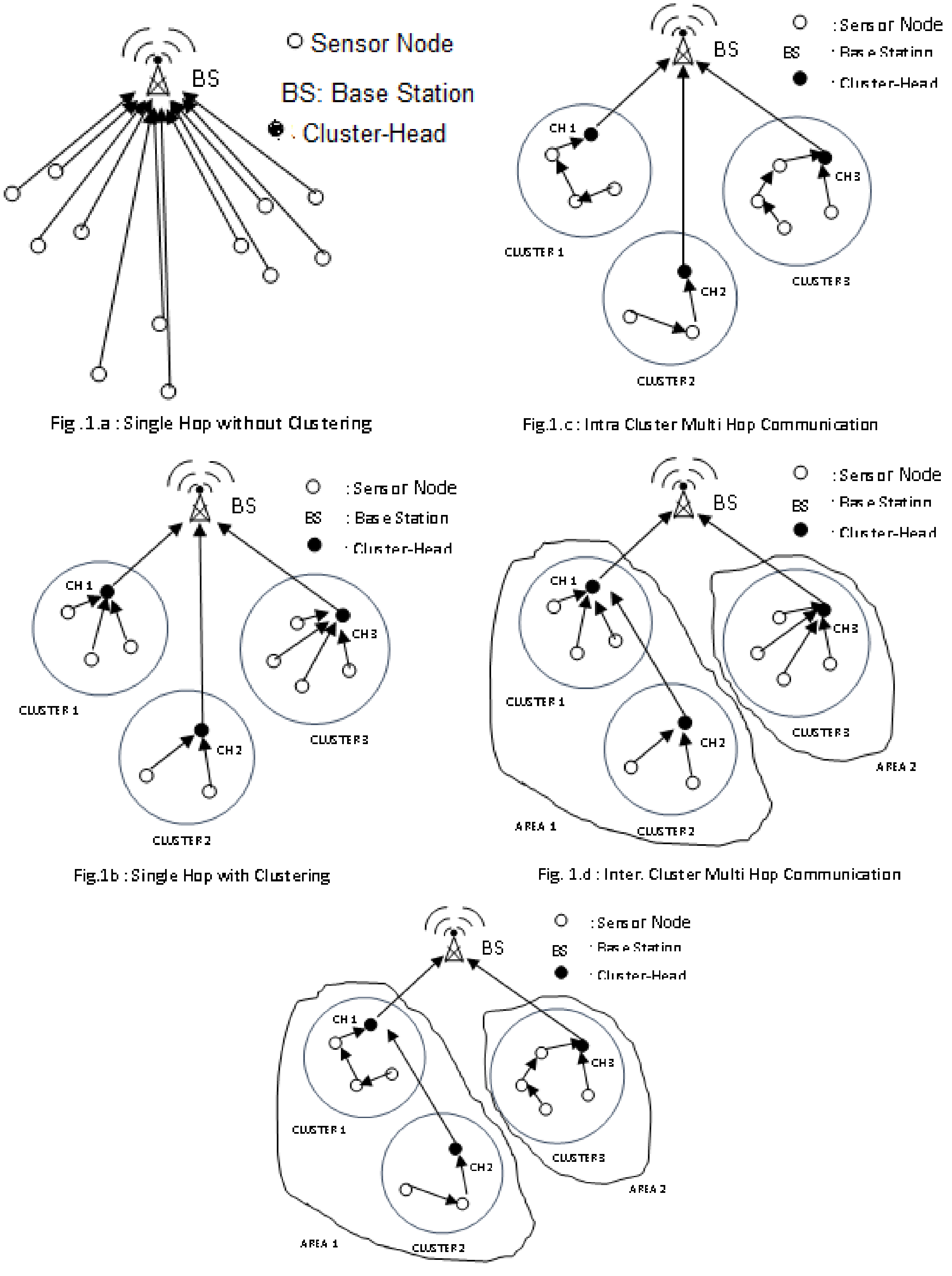


Fig .1.a : Single Hop without Clustering

Fig.1.c : Intra Cluster Multi Hop Communication

Fig.1b : Single Hop with Clustering

Fig. 1.d : Inter. Cluster Multi Hop Communication

Fig. 1.e : Both Intra. and Inter. Cluster Multi Hop Communication

Fig. 1: (a) Single Hop without Clustering (c) Intra Cluster Multi Hop Communication (b) Single Hop with Clustering (d) Inter. Cluster Multi Hop Communication (e) Both Intra. and Inter. Cluster Multi Hop Communication

to build Over lapping clusters (overlapping) to reinforce the robustness of the network and to respond to specific issues, namely the inter-cluster transmission,node localization and time synchronization.

The overlapping clusters construction in a distributed manner is an NP-hard problem; the protocol KOCA solves this problem in arandom and distributed way. KOCA,be declared regardless of complexity of the network size.

The overlapping clusters training problem is formulated by KOCA by building a set of cluster-Heads satisfies the three conditions of coverage,overlapping and connectivity

B. LMEEC (Layered Multi-Hop Energy Efficient Cluster-based)

Manel Khelifi and Assia Djabelkhir [3] proposed in 2012, MEC, a new multi-hop clustering protocol based on layered and energy efficiency, which offers a new way to reduce the energy consumption of sensors.

In order to provide flexibility for routing data through the network, LMEEC introduced a layered topology for the network nodes according to the number of hops that each of them takes to reach the base station. Thus, to achieve high energy efficiency and increase network scalability, sensor nodes are organized into clusters. To this end, they define a new, grouping mechanism node into clusters. This mechanism ensures distribution of the workload of sensor nodes by structuring them into unequal size clusters. Then the cluster-Heads communicate the data collected from the network to the base station. The clusters Head are periodically selected according to weight. This weight is calculated so that the number of Cluster-head increases approaching the base station[9].

Therefore, the further cluster of the base station have the smaller sizes.

The execution of LMEEC is periodically established in threephases.The first is the network configuration,while the second ensures the election of cluster-heads and the formation ofclusters.Data communication is the third phase of the protocol.

C. MCR (Multi-hop Clustering Routing Protocol)

S. Koteswara Rao, M. and T. SailajaMadhu [1], proposein 2012a protocol Clustering, called Multi-hop Clustering Routing Protocol(MCR), based on the use of Gate way nodes to achieve delivery data to are motebase station with are as onablecost in energy.

The MCR protocol uses a principle of inter-cluster transmission with two jumps,ie,CH snodesdo not communicate directly with the well,butthey usean intermediate node(Gateway) that is located in an area covered by the base station.For nodes not covered by the well,the protocol proposes to use the same principle as the HEED protocol for building clusters.

D. khLCH

The contribution of Khaled BOUCHAKOUR [4] in 2012 consists of a hierarchical routing protocol, called KhLCH (K.hop Layered Clutering Hierarchy), which aims minimizing energy consumption, scalability and reduced data delivery time. His solution uses K.Clusters formed on a restructured layered network; it allows for multi-hop communications, intra. and inter. Clusters, and collaborative data aggregated at Cluster-head and Gateway Nodes.

His solution is initially minimizing energy consumption and scalability; nodes are organized into layers according to their minimum distances, number of hops, the base station (the basic idea of LCH protocol). Then, these nodes are, organized into k clusters where each member node is either Cluster-headbe a member to k-hop CH (using a modified version of the KOCA algorithm). The organizing process can be divided into four (04) phases, Initialization, the Construction of k-clustering, data dissemination and maintenance of the topology.

E. Multihop-LEACH amlior

J S Rauthan and S. Mishra [5], in 2012 ,proposed multihop-LEACH improved protocol, which is one of the routing algorithms. The basic operation of multi hop-LEACH is similar toLEACHprotocol.There are two majorchanges in the multi hop-LEACH protocol compared to LEACH protocol. The multi-jump is applied in both the inter. and intra. Cluster communication.In this enhanced version of multi hop-LEACH protocol, the cluster contains; CH (responsible only to send the data that is received by the cluster members at the SB), vice-CH (the node becomes acluster of CH incase CH dies),cluster nodes(collect data from the environment and send to CH).

F. PUCMR (Partition Based Unequally Clustered Multi-Hop Routing Protocol)

U. Hari and Chris Johnson A[6], in 2013,state that the PUCMR protocol and unlike other uneven clusters based protocols, not only reduces the hot spot problem,but the issue of the unequal distribution of cluster-head is also eliminated. The proposed algorithm uses PUCMR energy,the degree of a node and the distance from center of gravity for the selection of cluster-Head and provides better position in the network.

Simulation results show that this approach extends the network lifetime.

G. Assisted-LEACH

Sunkara Vinodh Kumar and Ajit Pal,[7], in 2013, proposed the protocol Assisted-LEACH(A-LEACH) and declare that he has reached the level decreased and uniform distribution of the energy dissipated by the separation of routing tasks and aggregation of data.Heintroduced the concept of helper nodes(Nodes Helper) who assist Cluster Heads for Multi-hop routing.This algorithm has been, developed to facilitate energy efficiency,the configuration of the multi-hop route helper nodes to reach the base station.

III. Comparison and Synthesis

A. Comprative Table

The batch protocols that we have chosen, reports of recent work.Our comparison was, based on comparative factors shown in the table above.

TABLE I: Terminology

Symbole	Signification
CL	Clustering use
NC	Nature of generated clusters
EC	Cluster Election
RC	Cluster-Heads Rlection
ER	Consideration of node's Residualen ergy when selecting CH
CaC	Intra-Cluster Communication
CrC	Inter-Cluster Communication
ECH	Powers upporta large-scale network
GT	Use of Gatewaynodes
DA	Data Aggregation

TABLE II: Comparison

Protocol / Metrics	KOCA(2009)	LMEEC(2012)	MCR(2012)	khLCH(2012)	MultiHop-LEACH Improved(2012)	PUCMR(2013)	Assisted-LEACH(2013)
CL	Yes	Yes	Yes	Yes	Yes	Yes	Yes
NC	Static, Overlapping	-	dynamic, disjoint	Overlapping nodes	dynamic, disjoint	-	-
EC	random, probabilistic	random, probabilistic	random, probabilistic	random, probabilistic	random, probabilistic	random, probabilistic	random, probabilistic
RC	random, probabilistic	random, probabilistic	random, probabilistic	random, probabilistic	Vice-CH	random, probabilistic	random,probabilistic
ER	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CaC	No	No	No	Yes	Yes	No	Yes
CrC	Yes	-	Yes	Yes	Yes	Yes	Yes
GT	No	No	Yes	Yes	No	No	Yes
ECH	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DA	No	No	No	Yes	No	No	Yes

B. Synthesis

In the previous sections, we have established a state of the ar ton many study protocols;we developed a comparative table based on a number of comparative indicators. To this end, we selected khLC Hand Assisted-LEACH protocols as the most power ful because the yrespond favorably to strong majority of the criteria on which we supported this study

IV. The Study Protocol

A. LEACH (Low Energy Adaptive Clustering Hierarchy)

Definition

Heinzelman introduced a classification algorithm for sensor networks, called Low Energy Adaptive Hierarchical Clustering (Leach). LEACH is the first hierarchical cluster-based routing protocol for WSN. The advantage of this protocol is that it reduces the number of nodes that communicate directly with the base station and this by forming groups of cluster-heads. Then the other neighboring nodes connect and become members of that cluster, and consume a minimum of energy.

B. The operation mode of LEACH

The protocol takes place in rounds that have approximately the same pre-determined time interval. Each cycle begins with an initialization phase followed by transmission phase. The duration of the communication phase is longer than that of the construction phase (initialization) to minimize the overhead.

1) The Initialization Phase: The purpose of this phase is the construction of clusters by choosing leaders (CH) and setting the media to the access policy within each group

2) The Transmission Phase: This phase is longer than the previous one; it allows the collection of sensored data, using TDMA scheduler. The members transmit their data captured for their own slots; allowing them to turn off their communication interfaces outside their slots to save energy. This data is then aggregated by the CH that merges, compresses, and sends the final result to the base station.

After extinction of the CHs, the network will move to a new round. This process is repeated until all network nodes will be selected CH.

$$\tau(n) = \begin{cases} \frac{p}{1-p(r \quad mod \frac{1}{p})} & si \quad n \in G \\ 0 & sinon \end{cases} \qquad (1)$$

With:

- P: is the desired percentage of CH that is to say selected as $p = 0.05$ for optimal condition.
- r: current iteration.
- G: is the set of nodes that were not CHs at $(1/p)$ previous

C. Assisted-LEACH (Low Energy Adaptive Clustering Hierarchy)

Protocol Overview

The Assisted-LEACH Protocol sometimes called A-LEACH, includes the following sub-steps:o Selection of Cluster-Heads (CH)

- Creation of Clusters
- Selecting Helper Nodes
- The routing configuration
- Perception, aggregation and routing

D. The operating mode

In most clustering protocols, the entire load aggregation routing of data is done by the Cluster-Heads. The LEACH protocol transmits aggregated data directly by the cluster heads to the base station. This shortens the lifetime of the network. The concept of Helper Nodes was introduced, where a node closer to the base station in each cluster is assigned a routing task, while the cluster-Heads aggregate data. For the formulation of the route for helper nodes to reach the base station, each helper node is selected as the next hop; it is the nearest node to the base station among all neighboring helper nodes.

$$T(n) = \begin{cases} \frac{p}{1-p(r \quad mod\frac{1}{p})} & si \quad n \in G \\ 0.5\frac{p}{1-p(r \quad mod\frac{1}{p})} & si \quad n \in H \\ 0 & sinon \end{cases} \qquad (2)$$

With:

- P: desired percentage of Cluster-Head
- r: the current iteration in the operation protocol
- G: set of nodes that onttni-Cluster Head Nodes nor Helper in $[1/p]$ past iterations.
- H: set of nudsqui not been ontpas Cluster-Heads but who played the role of helper nodes in $[1/p]$ past iterations

V. Simulation and Results Interpetation

A. Presentation of NS2 Network Simulator

Network Simulator is a discrete event simulator for networks and is mainly used for the simulation of all levels of communication protocols, providing support for wired as well as wireless networks. It was designed in C ++ and provides a simulation interface through otcl, an object-oriented language Tcl. The user must describe the network topology by writing otcl scripts that are then executed by the NS-2 main program

TABLE III: Simulation Parameters

Parameters	Dimenssion
Area	100x100
Number of nodes	10
Nodes initial energy	2 joules
Percentage of Cluster	0.1
Energy consumed at data reception	W
Energy consumed at data transmission	W
Antenna Model	Omni Antenna
Type of Traffic	CBR
CBR Package Size	32 octets
Communication Model	Bi direction

B. Experience and Discussion

After completing our program ".tcl 'related to protocols and A-LEACH LEACH on NS2.35 we raised output the graphs below which and the respective interpretations are: He graph above represents the activity of the Task Manager of both LEACH and A-LEAH protocols

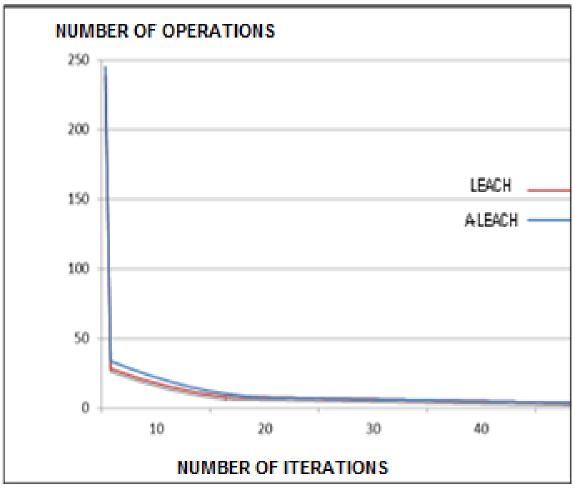


Fig. 2: Activity Task Manager

x-axis; we have the number of iterations.

In ordinate, we have the number of operations performed by the scheduler.

Sending messages is done on a sliding scale, hence the reason why at the start of network deployment, the task

manager took charge of time allocation to serve all the nodes, but as network nodes disappear during the iterations, the scheduler increasingly lightens its operations until vacancy with the expiry of all nodes.

The results obtained from the trace file of the two routing protocols (LEACH and A-LEACH) are almost identical. They explain why the Task Managers of both protocols start with a total of 225 operations.

Then the LEACH manager begins to decrease gradually until it reaches 2 through the 37th iteration. At the 38th iteration, it ends up with one and only one operation until the 51st iteration with network extinction, which means that at this stage it only manages itself.

The manager of A - LEACH begins to decrease gradually until it reaches 2 through the 39th iteration. And at the 40th iteration, it ends up with one and only one operation until the 53rd iteration with network extinction, which means that at this stage, as in LEACH, it only manages itself. The graph above illustrates the concept of traffic communication between the cluster nodes and the cluster Head At x-axis, we have the time with 103 as a time unit.

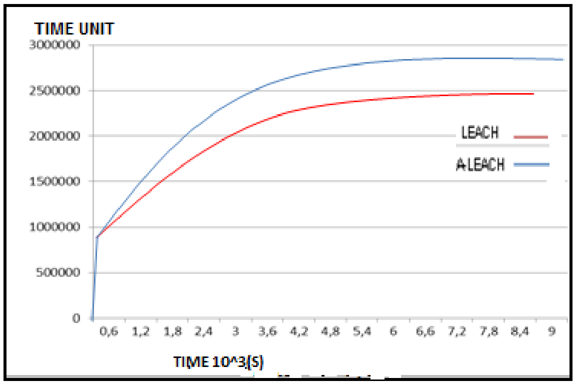


Fig. 3: the packages sent

In ordinate, we have the number of bits sent from the cluster and Head Node Helper. Sending messages is done progressively, hence the reason for the network stability in its early life and the disappearance of its nodes as a result of intensive messages ending. We collected figures from the trace file that shows for:

LEACH: 0 bits that start with the launch of the network, the number of packets sent from the cluster continues to rise, reaching 2.404 million bits sent after 780 time units, and remains stationary after explaining that there are more packet exchange due to the disappearance of all nodes.

A-LEACH: 0 bits that start with the launch of the network, the number of packets sent from the cluster continue to increase, to reach 2.668 million bits sent after 800 time units, and remains stationary until there is no more packet exchange due to the disappearance of all nodes.

The results obtained show that the packages sent by the cluster nodes using the A-leach protocol are growing compared to those sent by the nodes that use the LEACH protocol and this is because all nodes using protocol A-leach pass in a standby mode to reduce energy consumption.

The above graph illustrates the notion of life of nodes and that of the network. At x-axis, we have time.

In ordinate we have the number of operational nodes.

Our graph and the results retrieved from the log file on the network life reflect the following results:

For Leach: results show network stability on energy consumption and preserving all of its nodes for almost 0.88 * 103 s after which the network starts to record the gradual disappearance of its nodes to total extinction time in an approximate of 1.02 * 103s.

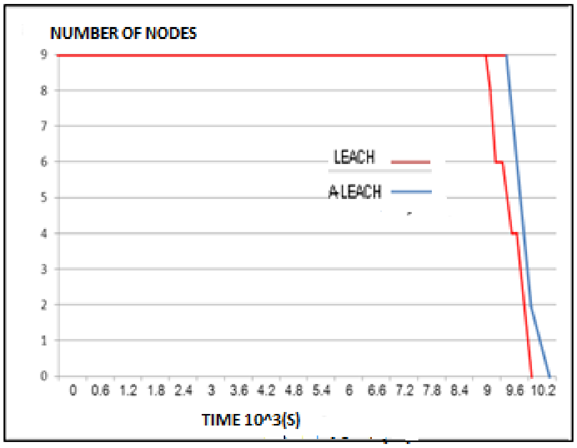


Fig. 4: lifetime

For A-leach: results shows network stability on energy consumption and preserving all of its nodes for almost 0.96 * 103 s after which the network starts to record the gradual disappearance of its nodes to total extinction in an the approximate time of 1.06 * 103s.

Consequently the life of a network using the protocol A-Leach is longer compared to a network using the Leach protocol.

C. Synthesis

Based on simulation results, we have shown that the A-LEACH protocol improves energy dissipation within the clusters, increases energy gain, and extends the network lifetime compared to LEACH protocol.

So the A- LEACH protocol provided the best value because it increases network lifetime

VI. Conclusion

Research in the field of sensor networks.

Is in full swing. Several routing protocols have been.

Developed in recent years. In this article we revised some routing protocols, with the aim of making a study of the performance of the latter.

We thought it useful to give an overview of the parameters (Metrics) used in the literature. We have extracted the better measures to measure performance in terms of Loss of packets and to decide the best of them under Conditions.

The work we have done (simulation under NS-2),

Allowed us to see the impact of many Nodes (or density), the energy consumed by the nodes and the variation of the scale, the rate of loss for protocols LEACH and A-LEACH

REFERENCES

[1] S. Koteswararao, M. Sailaja et T. Madhu, ”Implementation of Multi-hop Cluster based Routing Protocol for Wireless Sensor Networks”, International Journal of Computer Applications (0975 - 8887)Volume 59- No.8, pp 2-5, 2012.

[2] Moustafa A. YOUSSEF, Adel YOUSSEF, Mohamed F. YOUNIS, ”Overlapping MultihopClustering for Wireless Sensor Networks”, IEEE Transactions On Parallel And DistributedSystems, Vol. 20, No. 12, 2009.

[3] ManelKhelifi, AssiaDjabelkhir, ”LMEEC: Layered Multi-Hop Energy Efficient Cluster-basedRouting Protocol for Wireless Sensor Networks”, ReSyD, Doctoral School in Computer Science UAMB, Bejaia university, Algeria IEEE Transactions On Parallel And DistributedSystems,pp 1-2,2012.

[4] Khaled BOUCHAKOUR, ”Routage hirarchique sur les rseaux de capteurs sans fil: Protocole KhLCH (K-hop LayeredClusteringHierarchy)”, MEMOIRE Prsent pour l’obtention d’un diplme de MAGISTER EN INFORMATIQUE, pp 2-5, 2012.

[5] J. S. Rauthan, S. Mishra, ”An improved Cluster Based Multi-hop Routing in Self-Organizing Wireless Sensor Networks”, International Journal of Engineering Research Technology (IJERT) Vol.1, Issue 4, June - 2012.

[6] U. Hari, Chris Johnson A” A Partition BasedUnequallyClustered Multi-Hop Routing Protocol forWireless Sensor Networks”, International Journal of Engineering ResearchTechnology (IJERT) Vol. 2 Issue 5, pp1714-1718, 2013

[7] SunkaraVinodh Kumar, Ajit Pal, ” Assisted-Leach (A-Leach)Energy Efficient Routing Protocol for Wireless SensorNetworks”, International Journal of Computer and Communication Engineering, Vol. 2, No. 4, pp 420-424,2013.

[8] Herman D. Hughes,” Adaptive QoS Routing by Cross-Layer Cooperation in Ad Hoc Networks,” EURASIP Journal on Wireless Communications and Networking”, pp 661-671, 2005

[9] Draves, Richard and Padhye, Jitendra and Zill, Brian, ”Routing in Multi-radio, Multi-hop Wireless Mesh Networks”, Proceedings of the 10th Annual International Conference on Mobile Computing and Networking, vol MobiCom ’04, pp 114-128, 2004.

A Low Cost FPGA based Cryptosystem Design for High Throughput Area Ratio

Muhammad Sohail Ibrahim*, Irfan Ahmed†, M. Imran Aslam†, Muhammad Ghazaal*,
Muhammad Usman*, Kamran Raza* and Shujaat Khan*

*Faculty of Engineering Science and Technology (FEST),
Iqra University, Defence View,
Karachi-75500, Pakistan
†Department of Electronic Engineering,
NED University of Engineering and Technology,
University Road, Karachi 75270, Pakistan

Abstract—**Over many years, Field Programmable Gated Arrays (FPGA) have been used as a target device for various prototyping and cryptographic algorithm applications. Due to the parallel architecture of FPGAs, the flexibility of cryptographic algorithms can be exploited to achieve high throughputs at the expense of very low chip area. In this research, we propose a low cost FPGA based cryptosystem named as Secure Cipher for high throughput to area ratio. The proposed Secure Cipher is implemented using full loop unroll technique in order to exploit the parallelism of the proposed algorithm. The proposed cryptosystem implementation achieved a throughput of 4600Mbps for encryption. The logic resource utilization of this implementation is 802 logic elements(LE) which yields a throughput to area ratio of 5.735Mbps/LE.**

Keywords—Encryption; Cryptosystem; Secure Cipher; AES; FPGA; Full loop unroll

I. Introduction

Data security has been a topic of major interest since decades. With the development of communication systems, the techniques of data exchange have been revolutionized hence the need of data integrity and authenticity has also elevated. Various cryptosystems have been proposed in this regard. A cryptosystem is a software or a hardware that can convert data from its original comprehensible form into a scrambled form in such a way that the original information can be disclosed to some selected persons only [1], [2], [3]. Cryptosystems have evolved over the years from Ceaser's cipher, which was based on just shifting of letters, to the modern AES (Advanced Encryption Standard) proposed by Joan Daemen and Vincent Rijmen[4].

Cryptographic hardware solutions have been yet another field of interest for many researchers [5], [6]. Various hardware cryptosystems have been proposed in which the choice of hardware may be microcontrollers, microprocessors, and custom ASICs based cryptosystems. Each of the aforementioned hardware offer some merits and demerits, for instance, a microcontroller based design might have low processing capability but such a design usually takes low time to market. Similarly, an ASIC based solution can achieve very high data rates and power efficiency but require high time to market.

The hardware based designs can be compared on the basis of the following performance metrics. Power consumption, time to market, and Non Recurring Engineering (NRE) cost etc. Microcontroller based designs can be a choice for hardware implementation of cryptosystems as these designs are low cost and low power solutions and require very low time to market but their performance is also very low. For high performance requirements, a microprocessor based solution can be opted but such designs run on high power and their cost is also very high. Another class of microprocessor based solutions offer low cost and low power designs, but such microprocessors based solutions also offer very low performance. Hardware based solutions with high performance and low power can be designed on custom ASIC platform. ASIC designs are usually produced in mass volumes, so their per unit cost is also low but these solutions have high time to market as the generation of ASIC designs is a very complex process and in case of any error in the design the ASIC solution is redesigned which increases the NRE cost. For a high performance solution with low cost and low power consumption, FPGA based design is another candidate. These designs have very low time to market and have very low NRE cost of FPGAs due to the reconfigurability. The speed and efficiency of FPGAs combined with their flexibility makes them very attractive for cryptographic applications. The ability to reconfigure an FPGA to use a different cryptographic algorithm on the fly or to be able to update, modify or even replace an outdated algorithm make them very useful for designing cryptosystems. Likewise, low power and subsequently high throughputs that FPGAs are capable of make them very useful in high speed communications links or servers that often require security.

A. FPGA based Cryptosystem

There have been many FPGA based cryptosystem designs which focused on obtaining high throughputs. These designs often fully unroll the iterative round structure of the cryptosystem and rely heavily on pipelining within each round to increase throughput. High throughput FPGA designs typically achieve throughput above 20 Gbps and are intended to use in solutions that need to handle multiple security sessions simultaneously.

An FPGA based implementation of AES proposed by

T. Hoang used an iterative looping technique to implement AES for a block size of 128-bits[7]. In [8] another compact implementation of AES on FPGA is proposed. AES with block size of 128-bits was targeted to be implemented on FPGA. The key objective of that implementation was to keep the design as small as possible. The design achieved a throughput of 166Mbps at the expense of 222 slices and 3 block RAMs of 4Kbits each. In [9], the design decisions that lead to area/delay trade-offs in a single chip FPGA based cryptosystem is explored for AES. The design achieved a throughput of 23.57Gbps with 16938 slices of hardware area. G. Rouvroy proposed an efficient solution to combine AES encryption and decryption in one FPGA design keeping focus on low area constraint[10]. The proposed design achieved a throughput of 208Mbps using 163 slices and 3 blocks RAM only. In another research[11], a high performance encryptor/decryptor core of AES is presented. The design was implemented on a single-chip FPGA using fully pipelined technique. It uses 5677 slices and resulted in 4121Mbps throughput. Similarly, in [12], a fully pipelined AES encryption only design is presented. The design implemented on a single FPGA chip achieved a throughput of 21.54Gbps using 84 block RAMs and 5177 slices. In [13], another low power and low cost hardware core of AES algorithm is proposed. The core was designed with a novel 8-bit architecture that supports encryption with a 128-bit key. The design produces 121Mbps throughput at 153MHz clock frequency. In [14], four different architectures for AES-128 bits algorithm implementation are proposed. The four design techniques proposed in [14] are accurate floor-planning, unrolling, pipelining and tiling. These architectures were derived for different area-delay trade-offs. In [15], an efficient pipelined hardware implementation of AES-128 is proposed. The implementation will stay efficient even after increasing the required number of rounds to encounter attacks.The iterative looping with multi-stage sub-pipelining AES architecture is proposed in [16]. The design achieved 1.33Gbps throughput at 425MHz operating frequency. The logic resource utilization of the design is 303 slices. Another low cost AES implementation was proposed in [17]. This implementation proposed a high throughput design by the introduction of parallel operation in folded architecture. This implementation produced 37.1Gbps throughput at the maximum operating frequency of 505.5MHz.

Besides the AES, various other algorithms are also used to design FPGA based cryptosystems. S. Singh recently proposed a hardware implementation of RSA algorithm[18]. The authors have implemented RSA encryption using left to right radix-2 montmgomery multiplier on Xilinix Spartan-3 device. The design had a logic area utilization of 503 slices. The RSA algorithm FPGA implementation achieved 79.546MHz maximum clock frequency. In [19], an encryption scheme for real-time video streaming and its FPGA implementation has been proposed.

The demand of lightweight cryptographic algorithms has greatly increased due to the development and use of low resource devices for communication. In [20] a lightweight cipher named HIGHT, that provides adequate security at limited resource utilization is proposed along with its FPGA implementation. The authors presented pipelined and scalar (LUT) implementations of HIGHT with a claim of 18 times improved throughput at 60% less power consumption in pipelined design as compared to their LUT based design. In [21], the Minalpher algorithm and its implementation on various FPGA devices with simple and pipelined architecture is proposed. The performance of Minalpher algorithm was evaluated on resource constrained hardware.

The encryption process in standard algorithms is usually carried out by creating confusion and diffusion in the data. This objective is achieved by various operations such as shifting, transposition, various logical operation, and multiplication operations. Modern advancements in the field of data security suggest the use of algorithms that can be embedded in resource constrained devices such as smart phones, PDAs, etc [22]. Such devices have low on-board resources of memory and chip area, therefore, it is suggested to use algorithms with as low as possible complexity with adequate security. For this purpose, many researchers have proposed lightweight ciphers. The hardware implementation of such a lightweight block cipher named LEA is proposed in [23]. The algorithm was generally intended for software efficiency, therefore, the S-BOX structure was designed to have simple addition, rotation and XOR operations. The authors proposed a custom ASIC design which achieved a throughput of 533.3, 457.1, and 400 Kbps for key sizes of 128, 196, and 256 bits respectively at the operating frequency of 100KHz only. Furthermore, the design achieved 800Mbps throughput at 100MHz operating frequency for the key size of 256 bits. A full loop unroll architecture based FPGA implementation of a lightweight cryptographic algorithm named Secure Force is presented in [24]. The design achieved a throughput of 3.43Gbps at 53.5MHz operating frequency. In [25], an algorithm named Triple Hill Cipher, that can secure any binary data such as video, images, or audio data is proposed. The FPGA implementation of the algorithm achieved the maximum operating frequency of 528MHz at the expense of 4636 slices only.

B. Motivation and Organization of Paper

The ability of an FPGA to process data in parallel has attracted many researchers to use FPGA as a target device for the implementation and prototyping of a cryptosystem. Apart from keeping the algorithm efficient and lightweight, many programming techniques can be adopted to achieve high throughputs while keeping the chip area to the minimum. Such techniques include pipelining, full loop unrolling, sub-pipelining, partial loop unrolling etc [26].

In this paper, we propose a novel cryptosystem named Secure Cipher and its FPGA implementation. The rest of the paper is organized as follows; in section II, the proposed algorithm and its implementation is discussed. The experimental setup, evaluation criteria, and results are discussed in section III followed by the conclusion in section IV.

II. Proposed Cryptosystem and FPGA Implementation

The primary goals of any hardware cryptographic implementation are high throughput, low latency, low chip area, high operating frequency, and low power dissipation [27]. Since all these goals can never be achieved in a single hardware implementation, therefore, trade-offs are generally considered . These trade-offs are generally between delay or latency and chip area or resource utilization.

A. Secure Cipher

Many lightweight encryption algorithms have been proposed that are computationally inexpensive [28] The proposed Secure Cipher is low complexity encryption algorithm based on Feistal structure. It is a block cipher that consists of 5 encryption rounds only. Each encryption round consists of five logical and mathematical operations that operate on 8-bit data. This creates adequate confusion and diffusion in the data to confront various types of attacks. The proposed cryptosystem consists of the following blocks.

1) Key Generation Block: Key generation block generates five keys for each encryption and decryption round. The key generation block takes a 128-bit key as an input and generates round keys(K_r) of size 32 bits for each encryption/decryption round. The key generation block performs logical operations such as XOR and XNOR, fixed matrix multiplication, and left shift. Each of the logical notations have been displayed in figure I.

TABLE I: Notations and their Functions

Operation	Multiplication	XOR	XNOR
Notation	⊗	⊕	⊙

The input key (K) is an array of 128-bits, which is divided into 4 halves of 32-bits each. Each block of 32-bits is arranged in the form of a 4×8 matrix. Shift row operation is applied to each of the 4 matrices. Each of the shifted matrices are then arranged in an 4×8 matrix column-wise, on which XNOR logical operations are performed. The results of XNOR operation are stored in 4 matrices of the size 4×8 in column-wise fashion. These matrices then undergo a shift row operation and then multiplied with 4 individual fixed matrices of the size 8×4. The four fixed matrices labelled FM_1, FM_2, FM_3, and FM_4 are defined in equations (1),(2),(3), and (4) respectively. The detailed diagram of key generation block is shown in figure 1.

$$FM_1 = \begin{bmatrix} 128 & 64 & 32 & 16 & 8 & 4 & 2 & 1 \\ 64 & 32 & 16 & 8 & 4 & 2 & 1 & 128 \\ 32 & 16 & 8 & 4 & 2 & 1 & 128 & 64 \\ 16 & 8 & 4 & 2 & 1 & 128 & 64 & 32 \end{bmatrix} \quad (1)$$

$$FM_2 = \begin{bmatrix} 8 & 4 & 2 & 1 & 128 & 64 & 32 & 16 \\ 4 & 2 & 1 & 128 & 64 & 32 & 16 & 8 \\ 2 & 1 & 128 & 64 & 32 & 16 & 8 & 4 \\ 1 & 128 & 64 & 32 & 16 & 8 & 4 & 2 \end{bmatrix} \quad (2)$$

$$FM_3 = \begin{bmatrix} 128 & 32 & 64 & 8 & 16 & 1 & 4 & 2 \\ 64 & 128 & 8 & 1 & 32 & 4 & 2 & 16 \\ 1 & 16 & 4 & 32 & 128 & 8 & 64 & 2 \\ 32 & 2 & 128 & 4 & 16 & 64 & 1 & 8 \end{bmatrix} \quad (3)$$

$$FM_4 = \begin{bmatrix} 2 & 16 & 64 & 128 & 1 & 32 & 4 & 8 \\ 64 & 1 & 4 & 16 & 32 & 128 & 16 & 2 \\ 1 & 128 & 32 & 16 & 4 & 2 & 64 & 8 \\ 4 & 1 & 128 & 32 & 64 & 8 & 16 & 2 \end{bmatrix} \quad (4)$$

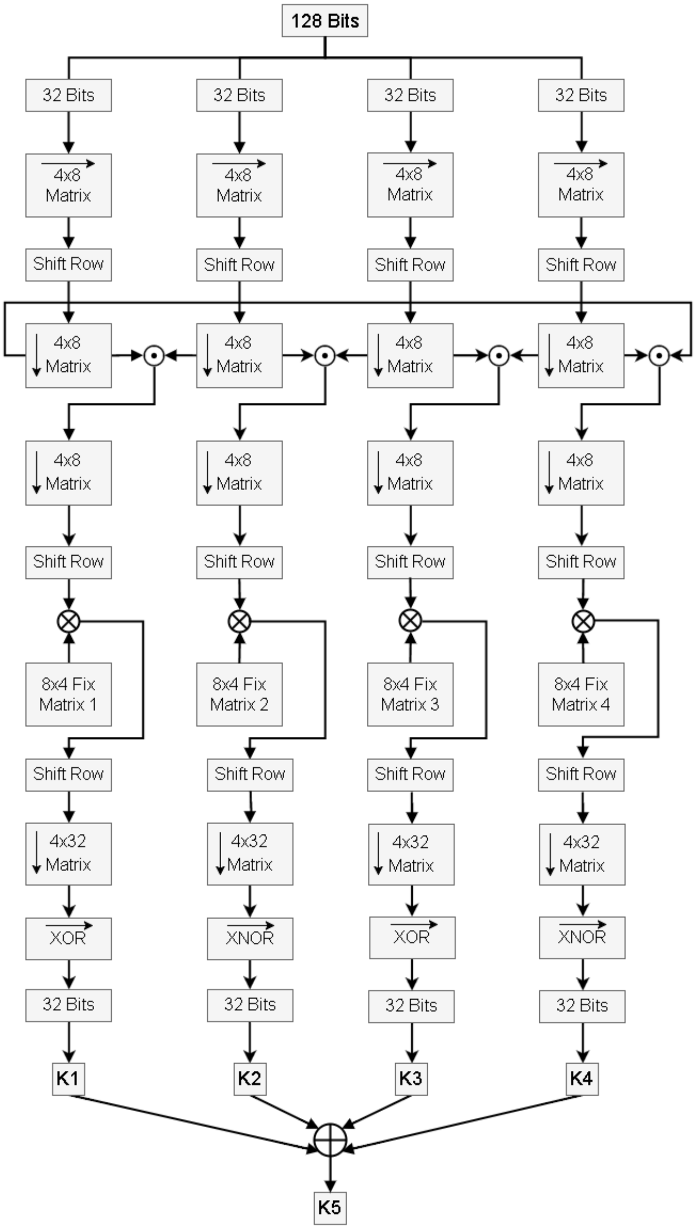


Fig. 1: Key expansion

2) Encryption Block: The encryption process consists of very simple operations of XOR, XNOR, left shift (LS), swapping operations, and Substitution Boxes (SBOX). The 128-bits wide plain text (X) is divided into two parts of 64-bits each, and these 64-bit halves are further divided into 32-bits. Swapping of 32-bits is performed in each round in order to alter the position of data hence increasing the complexity of the cipher. The round keys (K_r) are XNOR with the data in each round as shown in figure 2.

The block labelled ”F” in figure 2 is the principle block in encryption process as it contains the substitution boxes. Figure 3 presents the process of the F function. The 32 bits input of F block are divided into 4 halves of 8-bits each. The first 8

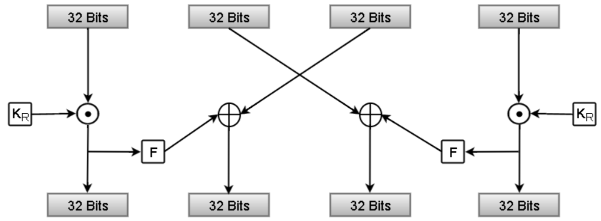


Fig. 2: A single encryption round

bits are moved to SBOX1 without performing any left shift operation. The second, third, and fourth 8 bit halves are left shifted by 1, 2, and 3 bits respectively. These left shifted halves are then moved to moved to the substitution boxes (SBOX1, SBOX2, SBOX3, and SBOX4) and then the results of each SBOX is concatenated to form 32 bits again. The structure of each of the SBOX is shown in figure 4.

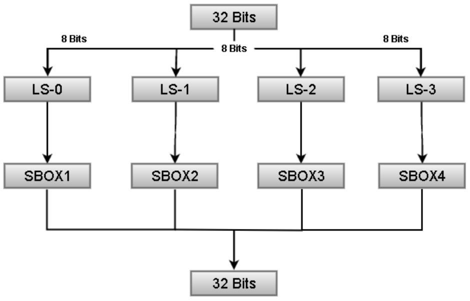


Fig. 3: F Function

An example of the substitution of data using SBOX is shown in figure 5. Each of the substitution box is generated in such a way that the output of any two or more than two substitution boxes cannot be the same despite the chance of having exactly the same selection byte. The SBOX operation takes place when the result of 32 bits data and round key K_R is divided into 4 halves of 8 bits each. Each of these 8 bit halves go through left shift operation as shown in figure 3. The 8 bit data is then moved to substitution boxes as the selection byte for the respective SBOX. The SBOX transformation takes place as; the 2 bits from MSB and 2 bits from LSB of the selection byte concatenate to give the row number of the SBOX, and the remaining 4 bits make the column number. In the example shown in figure 5, the output of the SBOX is $8C$, which is the $SB1_{(0,15)}$ entity of the SBOX1.

B. FPGA implementation

The overall hardware architecture of the Secure Cipher is based on loop unrolling technique. It is reported in [29] that loop unrolling is the main technique to achieve higher degrees of parallelism in reconfigurable hardware such as FPGA. It is also reported that loop unrolling increases the area but can also improve the throughput. The Secure Cipher is implemented on Altera Cyclone II EP2C35F672C6N FPGA using Verilog HDL.

As stated earlier that the design was lead out using full loop unroll technique. In this implementation, the iterations of

SB1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	63	7C	77	7B	F2	6B	6F	C5	30	O1	67	2B	FE	D7	AB	76
1	CA	82	C9	7D	FA	59	47	F0	AD	D4	A2	AF	9C	A4	72	C0
2	B7	FD	93	26	36	3F	F7	CC	34	A5	E5	F1	71	D8	31	15
3	O4	C7	23	C3	18	96	O5	9A	O7	12	80	E2	EB	27	B2	75
4	O9	83	2C	1A	1B	6E	5A	A0	52	3B	D6	B3	29	E3	2F	84
5	53	D1	OO	ED	20	FC	B1	5B	6A	CB	BE	39	4A	4C	58	CF
6	D0	EF	AA	FB	43	4D	33	85	45	F9	O2	7F	50	3C	9F	A8
7	51	A3	40	8F	92	9D	38	F5	BC	B6	DA	21	10	FF	F3	D2
8	CD	0C	13	EC	5F	97	44	17	C4	A7	7E	3D	64	5D	19	73
9	60	81	4F	DC	22	2A	90	88	46	EE	B8	14	DE	5E	0B	DB
10	E0	32	3A	0A	49	O6	24	5C	C2	D3	AC	62	91	95	E4	79
11	E7	C8	37	6D	8D	D5	4E	A9	6C	56	F4	EA	65	7A	AE	O8
12	BA	78	25	2E	1C	A6	B4	C6	E8	DD	74	1F	4B	BD	8B	8A
13	70	3E	B5	66	48	O3	F6	0E	61	35	57	B9	86	C1	1D	9E
14	E1	F8	98	11	69	D9	8E	94	9B	1E	87	E9	CE	55	28	DF
15	8C	A1	89	0D	BF	E6	42	68	41	99	2D	0F	B0	54	BB	16

Substitution box # 1

SB2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	O4	93	8O	C9	24	BC	5F	88	E5	47	OF	81	5E	9O	6D	34
1	75	38	4C	OB	1C	8O	8E	DB	92	74	B4	C4	2D	E9	A6	F1
2	OC	D8	O7	B3	3A	6A	O7	4C	FO	C5	77	68	F4	B7	6D	5D
3	D9	36	CC	87	81	26	E5	27	B7	96	B2	6B	E1	92	D6	1O
4	9B	AA	3F	CO	1D	EO	89	92	6O	F3	OE	41	9B	D1	BE	EC
5	AC	A5	73	9O	93	69	8O	D5	A1	65	23	AF	81	E5	C5	32
6	72	FE	BD	5B	CD	BF	3D	53	DE	E7	72	21	8F	B3	B1	O5
7	2A	4E	6F	OF	5E	62	BF	AO	61	36	6D	49	62	89	AB	E8
8	74	67	AF	2A	78	A8	FC	A7	E5	24	OF	O5	59	7B	6F	EC
9	8F	EA	25	D5	7C	DA	78	34	54	BO	EB	A9	3O	ED	6O	35
10	CD	38	6O	68	6O	47	AD	6O	B2	41	8A	DE	CF	75	88	C7
11	36	EC	12	24	5C	CC	C2	A1	F6	AC	BB	EF	C2	28	7A	13
12	B9	CA	5F	35	8O	3O	71	34	54	25	75	E5	A1	66	44	3C
13	D4	57	D7	FF	OB	74	DF	F2	37	E8	8O	CA	A6	D4	C8	3F
14	9B	32	69	3A	E3	DA	B2	BO	DC	13	7O	5O	AO	98	9E	B6
15	4A	E6	2O	24	B8	42	C5	1F	8A	6B	2F	CB	8A	95	88	91

Substitution box # 2

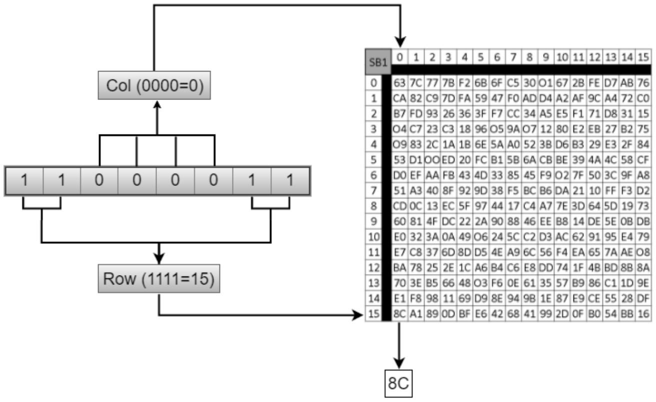
SB3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	82	89	7F	11	A5	4A	41	41	75	64	42	CC	14	83	6C	6D
1	A5	1B	A9	22	C8	5F	C4	9F	B2	FB	89	1D	D5	E1	8O	12
2	15	4C	E2	55	12	79	O7	2C	B8	OO	39	D7	7E	A1	13	49
3	88	48	B5	F1	72	8F	5O	28	O7	71	E4	C1	53	OA	27	79
4	EE	C8	79	88	77	24	44	DB	65	3O	73	7B	DE	14	23	85
5	DD	35	B6	51	86	D7	1E	B3	21	89	EB	BC	31	34	AB	C2
6	1F	58	OD	FA	B4	A4	97	46	OC	BF	96	9A	7B	D3	DB	A4
7	23	A5	36	76	69	52	E2	1D	DE	2F	53	EC	E6	F7	54	29
8	BA	7D	59	4E	4E	15	B3	O7	A9	AF	C7	A3	E4	DE	3A	1A
9	97	27	D9	O1	35	9B	19	1B	49	74	DB	7C	D6	3O	ED	FE
10	B1	29	D1	EO	86	FO	5D	61	A9	B3	C2	CO	FF	53	FE	1D
11	9B	A3	3A	EE	59	92	48	92	26	A8	B3	CE	F6	BO	45	59
12	49	7F	2C	AA	C1	75	AA	6D	C7	41	EF	DO	C4	DD	95	7D
13	1O	A3	43	C3	BC	C4	88	OO	AE	8D	72	D4	45	41	2E	39
14	9D	5B	EO	8O	AB	38	AD	D1	22	48	15	A6	65	C1	37	16
15	D4	5C	45	82	8O	3B	C6	5O	O5	56	DC	FO	68	BD	86	6F

Substitution box # 3

SB4	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	96	9D	79	6E	11	77	9F	67	BB	EF	1O	37	17	98	1A	D8
1	A6	4B	OF	75	3C	B2	CE	D8	61	29	15	56	8E	A3	F8	45
2	84	94	OD	BO	A1	4A	41	61	E5	59	D7	DE	F1	5E	D1	O9
3	7O	B4	87	AF	C9	AC	DA	O2	OF	7D	1C	2C	F9	45	O2	FB
4	DE	2E	31	2C	69	F8	CB	A3	B1	7C	6D	AO	14	EC	5B	C6
5	95	52	C2	6A	48	6A	71	BE	82	79	56	CD	2C	B8	21	DC
6	55	24	97	88	3A	CA	OB	34	91	EO	6D	98	44	11	FE	9D
7	57	25	98	C1	D3	54	CF	FC	5D	58	33	88	4O	16	O1	D9
8	8O	97	O3	FC	83	EF	O7	C8	3A	5A	OC	BC	D5	6D	C4	C2
9	C1	EB	55	36	A5	BD	B6	A2	1F	4D	45	A8	54	26	EF	7C
10	2C	15	37	1F	OD	62	88	46	85	21	FA	FO	19	B7	8C	68
11	D8	D3	69	3B	OD	BA	FE	7E	1O	25	BO	D8	BE	5D	34	D7
12	DO	41	8F	7B	D4	O8	C5	E7	31	14	17	2C	5C	5E	A7	9D
13	EO	F3	79	23	44	O2	OB	5C	6E	A7	5B	D5	CD	81	63	4C
14	B9	3E	7A	39	17	CA	O8	E1	O6	2B	E7	99	F7	F3	95	1C
15	E7	O4	31	DF	2F	3A	57	8A	99	FC	9D	AB	31	F7	8E	6F

Substitution box # 4

Fig. 4: Substitution Boxes



SB1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	63	7C	77	7B	F2	6B	6F	C5	30	O1	67	2B	FE	D7	AB	76
1	CA	82	C9	7D	FA	59	47	F0	AD	D4	A2	AF	9C	A4	72	C0
2	B7	FD	93	26	36	3F	F7	CC	34	A5	E5	F1	71	D8	31	15
3	O4	C7	23	C3	18	96	O5	9A	O7	12	80	E2	EB	27	B2	75
4	O9	83	2C	1A	1B	6E	5A	A0	52	3B	D6	B3	29	E3	2F	84
5	53	D1	OO	ED	20	FC	B1	5B	6A	CB	BE	39	4A	4C	58	CF
6	D0	EF	AA	FB	43	4D	33	85	45	F9	O2	7F	50	3C	9F	A8
7	51	A3	40	8F	92	9D	38	F5	BC	B6	DA	21	10	FF	F3	D2
8	CD	0C	13	EC	5F	97	44	17	C4	A7	7E	3D	64	5D	19	73
9	60	81	4F	DC	22	2A	90	88	46	EE	B8	14	DE	5E	0B	DB
10	E0	32	3A	0A	49	O6	24	5C	C2	D3	AC	62	91	95	E4	79
11	E7	C8	37	6D	8D	D5	4E	A9	6C	56	F4	EA	65	7A	AE	O8
12	BA	78	25	2E	1C	A6	B4	C6	E8	DD	74	1F	4B	BD	8B	8A
13	70	3E	B5	66	48	O3	F6	0E	61	35	57	B9	86	C1	1D	9E
14	E1	F8	98	11	69	D9	8E	94	9B	1E	87	E9	CE	55	28	DF
15	8C	A1	89	0D	BF	E6	42	68	41	99	2D	0F	B0	54	BB	16

Fig. 5: Sbox Transformation

every loop in the algorithm are unrolled in such a way that the output of each iteration becomes the input of the successive loop iteration. The hardware design of every sub-module of the Secure Cipher will be described in the respective subsections.

1) Key Genration Module: Key generation block generates the keys for individual rounds. This block takes 128 bits as an input key and performs various logical operations. This is to create enough confusion and diffusion in the input key in order to eliminate the chance of generation of weak keys. The key expansion process of the proposed Secure Cipher relies mainly on logical operations such as OR, AND, XOR, XNOR, Left shift, transposition, and permutation operations. Permutation and transposition operations are mapped by substitution.

Another operation included in key expansion block is the fixed matrix multiplication operation. There are four fixed matrices of the size 8×4, and these matrices hold fixed 8 bit integer values. As illustrated in figure 1, the output of the XNOR operation is arranged in an 4×8 matrix row-wise on which a left shift operation is applied. Each of these shifted matrices of 32 bits are multiplied with the fixed matrix which results into a 4×4 matrix of 128 bits. The obtained 4×4 matrix then goes through a left shift operation. We observed that the fixed matrix multiplication produces results from a finite set of numbers, as there involves multiplication of a binary 1 or 0 with an 8 bits wide number. Therefore, instead of using hard multiplier blocks, we transformed the fixed matrix multiplication problem into fixed look up tables. Each entity of the result of fixed matrix multiplication is defined by the equation shown as the output of the look up table presented in figure 6.

In figure 6, RS is defined as the row shifted matrix of size 4×8, FM is defined as the fixed matrix of size 8×4. The result of the equation is defined as the $(i,j)th$ entity of the matrix labelled as fixed matrix multiplication output FM_o.

The hardware of the fixed matrix multiplication is illustrated in figure 6. The select line of the look up table is the 8 bits wide row of the left shifted matrix RS, which depicts that the 8 bits wide output FM_o of the look up table can be selected form 256 possible input combinations. Each of the input is the product of ith element of the row of RS matrix with the jth element of the column of fixed matrix FM.

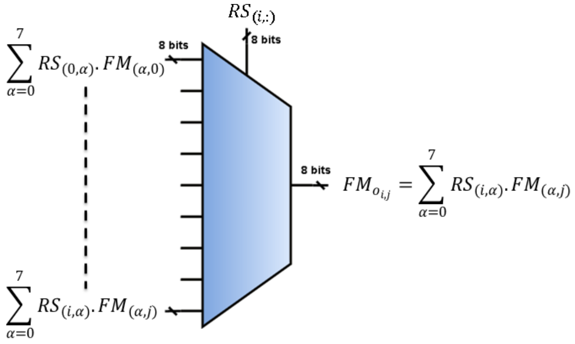

$RS_{(i,:)}$
8 bits
8 bits
$\sum_{\alpha=0}^{7} RS_{(0,\alpha)} \cdot FM_{(\alpha,0)}$
$\sum_{\alpha=0}^{7} RS_{(i,\alpha)} \cdot FM_{(\alpha,j)}$
8 bits
$FM_{o_{i,j}} = \sum_{\alpha=0}^{7} RS_{(i,\alpha)} \cdot FM_{(\alpha,j)}$

Fig. 6: Fixed Matrix Multiplication

2) Encryption Module: The encryption module of Secure Cipher consists of simple logical operations (AND, OR, XOR, and XNOR) and substitution boxes (SBOX). The encryption module takes 128 bits plain text as an input and divides it into 4 halves of 32 bits each. The encryption process continues as illustrated in figure 2. Since encryption is an iterative process, therefore full loop unroll technique is employed which unrolls all the five encryption rounds. The block diagram of loop unrolled encryption block is shown in figure 7.

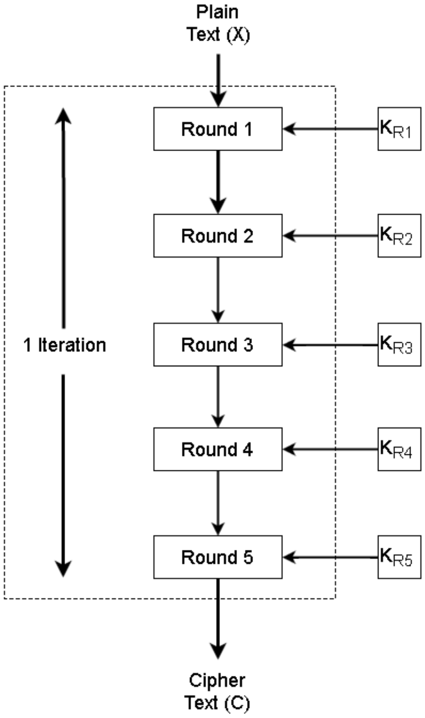


Fig. 7: Full Loop Unrolled Encryption

Each encryption round takes the output of the previous round and a round key K_R as an input. As mentioned earlier that the F function block displayed in figure 3 is the block of principle importance in encryption. Each of the SBOX in F function is an array of the size 16×16, which performs substitution. The hardware of SBOX, as shown in figure 8 is also a look up table which selects its output from 256 standard values. The selection of output is displayed in figure 5. The encryption process is the same in all of the five rounds. At the end of the $5th$ round, the 32 bits wide outputs are concatenated to form the cipher text or encrypted message.

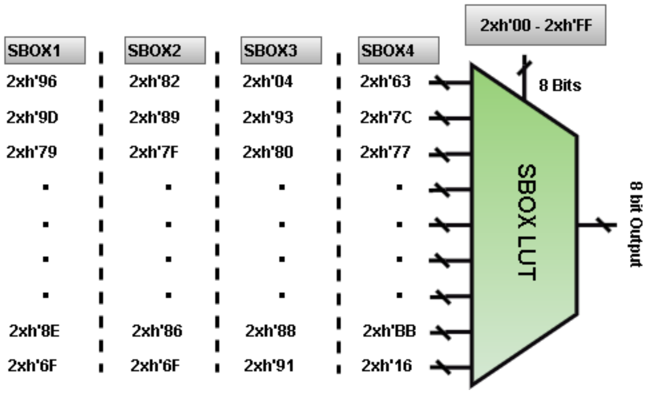


Fig. 8: Substitution box look up table

III. Experimental Setup

The security evaluation of cryptosystems is done on the well known parameters such as key senstivity test based on strict avalanche criterion(SAC), entropy, histogram, and correlation[30], [31], [32], [33], [34]. The hardware designs of cryptographic algorithms are generally compared on the basis of their logic resource utilization or area, propagation delay or latency, throughput, power consumption, and maximum operating frequency [35], [36], [37].

The target device for the proposed cryptosystem implementation is a low cost Altera Cyclone II EP2C35F672C6N FPGA. The details of the aforementioned evaluation parameters will be described in later subsections.

A. Evaluation Parameters

The performance of the Secure Cipher is evaluated on the following performance metrics. The results related to security were performed on MATLAB software. And the hardware performance evaluation parameters such as area, propagation delay, and throughput were performed on Altera Cyclone II FPGA using Quartus II 12.1 sp1 edition software.

1) *Key Sensitivity:* Key sensitivity of cryptosystems is tested on the basis of Strict Avalanche Criterion (SAC). The SAC states that ”If a function is to satisfy the strict avalanche criterion, then each of its output bits should change with a probability of one half whenever a single input bit is complemented”[38]. For key sensitivity test, the the cipher text should change with a probability of 50 %.

2) *Image Entropy and Correlation:* Entropy is the measure of information content of the data. The entropy of the encrypted data should be high so that the data cannot be recognized after encryption. And correlation is defined as the measure of similarity between the adjacent pixels of an image. For an efficient cryptosystem, the results of correlation of an encrypted image should be as low as possible so as to ensure that the data is scrambled adequately.

3) *Histogram:* For the security related testing, we performed the tests on image data since the results in the visual form can be understood easily. The histogram of an image before encryption shows the intensity variation of the image pixels. For an encrypted image, the pixel intensity should be uniform. This shows the randomness created in the image after encryption.

4) *Area:* The area in FPGAs is measured in terms of the logic units or circuits being used by the design. For Altera Cyclone II FPGA family, the resource utilization or area is measured in terms of the number of logic elements (LE), whereas for Xilinx Spartan FPGAs, the term logic circuits (LC) is used. A logic element (LE) contains a 4 input Look-Up Table (LUT), a D flip-flop, and a register for carry chain connection.

In [26], it is reported that the cryptosystems designed with full loop unroll technique may have larger area on hardware as compared to partial loop unrolled architectures, but such designs can achieve high throughputs.

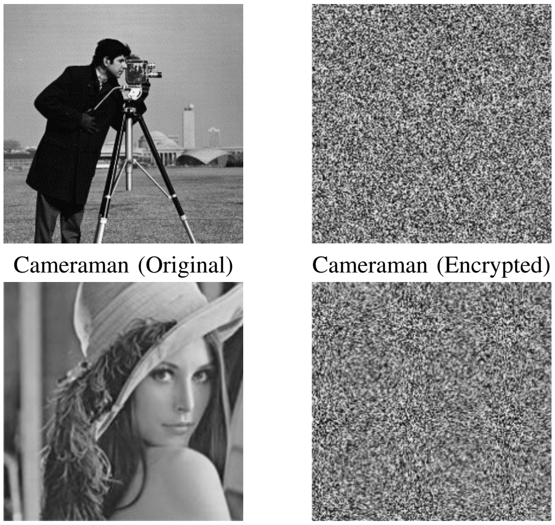
5) *Propagation Delay:* Propagation delay is defined as the maximum amount of time that exists between the edges of signal when it propagates from input to the output of a given circuit, so, it is the amount of time for the slowest signal to propagate from input to output in a circuit. The propagation delay can be greater if the circuit has complex operations and large area. In general, the propagation delay can be high for full loop unroll designs, but it can be low if the algorithm’s flexibility is properly utilized. For instance, in the proposed algorithm, the fixed matrix multiplication is the most complex mathematical operation and it can cause higher delays even if hardware multiplier blocks are used to perform multiplication. But instead of using the multiplier blocks, we propose to implement this multiplication on a simple look-up tables problem which is very low in terms of complexity as compared to the conventional multiplication operation. Such look-up tables implementations cause much less propagation delays as compared to hard multiplier blocks.

6) *Throughput:* Throughput is referred as the primary measure of speed for a hardware based cryptosystem. For hardware implementation of algorithms, throughput is the measure of the amount of data (in bits) processed per unit time. Modern hardware cryptosystems posses high speed data links, therefore, their throughputs should be high enough to be in orders of $Mbps$ to $Gbps$ so as to utilize the high data link speeds.

B. Results

The evaluation parameters related to security have been described in section III-A. The proposed Secure Cipher performs adequately in terms of security. The visual testing results of image encryption using the proposed Secure Cipher have been displayed in figure 9. The security related tests have been performed on images of the size 256×256 named Cameraman and Lena. It can be seen in the figure 9 that the encrypted images are impossible to identify visually.

The key sensitivity is tested on strict avalanche criterion (SAC). Based on the SAC, we calculated the mean percentage avalanche value for 1000 variations in the input key and plain text and achieved the mean percentage avalanche value of 54.55%.



Cameraman (Original) Cameraman (Encrypted)

Lena (Original) Lena (Encrypted)

Fig. 9: Image encryption visual results

As mentioned in section III-A, the histogram of the encrypted image should be uniform so that each pixel contains nearly the same information content. The histogram results of the original and encrypted images have been shown in figure 10. Whereas the correlation results of original and encrypted images have been shown in figure 11.

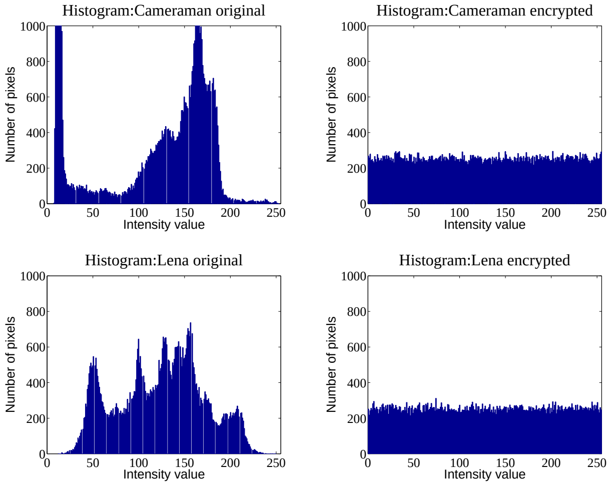


Fig. 10: Histogram of original vs. encrypted images

The proposed Secure Cipher was implemented on Altera DE2 board with Cyclone II EP2C35F672C6N FPGA. The design was synthesized using Quartus II 12.1 sp1 edition. The FPGA implementation results are listed in table II. In [26], it is reported that the hardware implementations with full loop unroll architectures may occupy high area. But the proposed Secure Cipher has low algorithmic complexity such that the resource utilization of the proposed Secure Cipher is lower than [15], [25], [23], and [39].

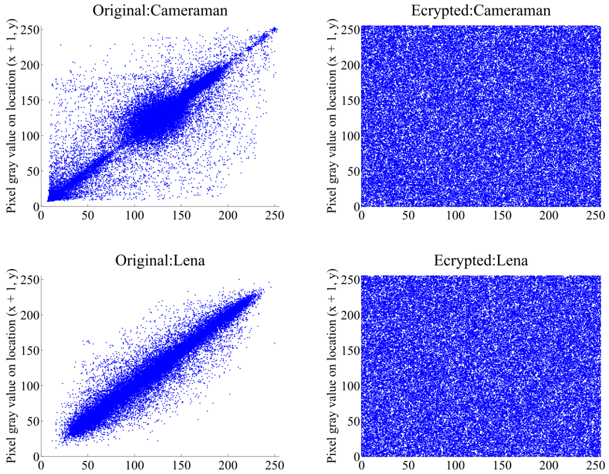


Fig. 11: Correlation of original vs. encrypted images

The throughput to area ratio should be high for a hardware cryptosystem as it shows the contribution of a single LE in the speed of the hardware design. It is evident from the results displayed in table II that the proposed Secure Cipher has higher throughput to area ratio than the designs presented in [15], [25], and [39] as mentioned in table II.

TABLE II: Comparison of Implementation Results

Design	**Device**	**Propagation Delay (ns)**	**Throughput (Mbps)**	**Area (LEs)**	**Throughput Area Ratio (Mbps/LE)**
HIGHT (2016) [20]	Cyclone II	14.97	4275.2	632	6.76
Triple Hill (2014) [25]	Virtex-4	1.894	67581.8	4636	14.57
LEA (2014) [23]	Cyclone III	200	650.19	813	0.8
DES (2015) [39]	Vertix II	2.182	278.26	303	0.918
Secure Cipher	Cyclone II	13.925	4600	802	5.735

IV. Conclusion

Reconfigurable hardware devices such as FPGAs play a vital role in assessing the performance of cryptographic block ciphers on hardware platform. The proposed cryptosystem named Secure Cipher was designed on FPGA using Full Loop Unroll architecture. The hardware performance results are promising in terms of area, and throughput as the complete design was implemented on Altera Cyclone II FPGA using 802 LE only. And the proposed system has a throughput of 4600Mbps with 5.735Mbps/LE throughput to area ratio. Whereas the proposed Secure Cipher ensures adequate security with a percentage SAC value of 54.55%. For future considerations, the pipelined design of the proposed cryptosystem can be implemented which would help in evaluating the flexibility of the proposed Secure Cipher.

References

[1] S. Khan, M. Ebrahim, and K. A. Khan, "Performance evaluation of secure force symmetric key algorithm," 2015.

[2] M. Ebrahim, S. Khan, and U. Khalid, "Security risk analysis in peer 2 peer system; an approach towards surmounting security challenges," *arXiv preprint arXiv:1404.5123*, 2014.

[3] M. Ebrahim, S. Khan, and S. S. U. H. Mohani, "Peer-to-peer network simulators: an analytical review," *arXiv preprint arXiv:1405.0400*, 2014.

[4] J. Daemen and V. Rijmen, "Aes proposal: Rijndael," 1999.

[5] S. Khan, M. S. Ibrahim, K. A. Khan, and M. Ebrahim, "Security analysis of secure force algorithm for wireless sensor networks," *arXiv preprint arXiv:1509.00981*, 2015.

[6] M. Ebrahim, S. Khan, and U. B. Khalid, "Symmetric algorithm survey: A comparative analysis," *International Journal of Computer Applications (0975 – 8887)*, vol. 61, no. 20, 2014.

[7] T. Hoang *et al.*, "An efficient fpga implementation of the advanced encryption standard algorithm," in *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on*. IEEE, 2012, pp. 1–4.

[8] P. Chodowiec and K. Gaj, "Very compact fpga implementation of the aes algorithm," in *Cryptographic Hardware and Embedded Systems-CHES 2003*. Springer, 2003, pp. 319–333.

[9] J. Zambreno, D. Nguyen, and A. Choudhary, "Exploring area/delay tradeoffs in an aes fpga implementation," in *Field Programmable Logic and Application*. Springer, 2004, pp. 575–585.

[10] G. Rouvroy, F.-X. Standaert, J.-J. Quisquater, and J.-D. Legat, "Compact and efficient encryption/decryption module for fpga implementation of the aes rijndael very well suited for small embedded applications," in *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, vol. 2. IEEE, 2004, pp. 583–587.

[11] F. Rodrıguez-Henrıquez, N. Saqib, and A. Dıaz-Perez, "4.2 gbit/s single-chip fpga implementation of aes algorithm," *Electr. Lett*, vol. 39, no. 15, pp. 1115–1116, 2003.

[12] A. Hodjat and I. Verbauwhede, "A 21.54 gbits/s fully pipelined aes processor on fpga," in *Field-Programmable Custom Computing Machines, 2004. FCCM 2004. 12th Annual IEEE Symposium on*. IEEE, 2004, pp. 308–309.

[13] P. Hämäläinen, T. Alho, M. Hännikäinen, and T. D. Hämäläinen, "Design and implementation of low-area and low-power aes encryption hardware core," in *Digital System Design: Architectures, Methods and Tools, 2006. DSD 2006. 9th EUROMICRO Conference on*. IEEE, 2006, pp. 577–583.

[14] G. P. Saggese, A. Mazzeo, N. Mazzocca, and A. G. Strollo, "An fpga-based performance analysis of the unrolling, tiling, and pipelining of the aes algorithm," in *Field Programmable Logic and Application*. Springer, 2003, pp. 292–302.

[15] N. Nedjah, L. de Macedo Mourelle, and C. Wang, "A parallel yet pipelined architecture for efficient implementation of the advanced encryption standard algorithm on reconfigurable hardware," *International Journal of Parallel Programming*, pp. 1–16, 2016.

[16] M. El Maraghy, S. Hesham, and M. A. Abd El Ghany, "Real-time efficient fpga implementation of aes algorithm," in *SOC Conference (SOCC), 2013 IEEE 26th International*. IEEE, 2013, pp. 203–208.

[17] K. Rahimunnisa, P. Karthigaikumar, S. Rasheed, J. Jayakumar, and S. SureshKumar, "Fpga implementation of aes algorithm for high throughput using folded parallel architecture," *Security and Communication Networks*, vol. 7, no. 11, pp. 2225–2236, 2014.

[18] S. Singh and P. S. Jassal, "Synthesis and analysis of 32-bit rsa algorithm using vhdl," 2016.

[19] F. Sbiaa, S. Kotel, M. Zeghid, R. Tourki, M. Machhout, and A. Baganne, "A format-compliant selective encryption scheme for real-time video streaming of the h. 264/avc," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 386–396, 2016.

[20] B. J. Mohd, T. Hayajneh, Z. A. Khalaf, A. Yousef, and K. Mustafa, "Modeling and optimization of the lightweight hight block cipher design with fpga implementation," *Security and Communication Networks*, 2016.

[21] M. Kosug, M. Yasuda, and A. Satoh, "Fpga implementation of authenticated encryption algorithm minalpher," in *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2015, pp. 572–576.

[22] D. Hong, J. Sung, S. Hong, J. Lim, S. Lee, B.-S. Koo, C. Lee, D. Chang, J. Lee, K. Jeong *et al.*, "Hight: A new block cipher suitable for low-resource device," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2006, pp. 46–59.

[23] D. Lee, D.-C. Kim, D. Kwon, and H. Kim, "Efficient hardware implementation of the lightweight block encryption algorithm lea," *Sensors*, vol. 14, no. 1, pp. 975–994, 2014.

[24] S. Khan, M. S. Ibrahim, H. Amjad, K. A. Khan, and M. Ebrahim, "Fpga implementation of 64 bit secure force algorithm using full loop-unroll architecture," in *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 2015, pp. 1–6.

[25] A. A. Khalaf, M. S. A. El-karim, and H. F. Hamed, "A triple hill cipher algorithm proposed to increase the security of encrypted binary dataand its implementation using fpga," *Journal Editorial Board*, vol. 1, no. 3, p. 752, 2014.

[26] A. J. Elbirt, W. Yip, B. Chetwynd, and C. Paar, "An fpga-based performance evaluation of the aes block cipher candidate algorithm finalists," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 9, no. 4, pp. 545–557, 2001.

[27] S. Khan, M. S. Ibrahim, M. Ebrahim, and H. Amjad, "Fpga implementation of secure force (64-bit) low complexity encryption algorithm," *International Journal of Computer Network and Information Security*, vol. 7, no. 12, p. 60, 2015.

[28] M. Usman, I. Ahmed, I. Aslam, S. Khan, and U. A. Shah, "Sit: A lightweight encryption algorithm for secure internet of things," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8(1), no. 51, 2017.

[29] B. Buyukkurt, Z. Guo, and W. A. Najjar, "Impact of loop unrolling on area, throughput and clock frequency in roccc: C to vhdl compiler for fpgas," in *Reconfigurable Computing: Architectures and Applications*. Springer, 2006, pp. 401–412.

[30] A. Kumar and M. N. Tiwari, "effective implementation and avalanche effect of aes," *International Journal of Security, Privacy and Trust Management (IJSPTM)*, vol. 1, no. 3/4, pp. 31–35, 2012.

[31] S. Shivkumar and G. Umamaheswari, "Performance comparison of advanced encryption standard (aes) and aes key dependent s-box-simulation using matlab," in *Process Automation, Control and Computing (PACC), 2011 International Conference on*. IEEE, 2011, pp. 1–6.

[32] M. Zeghid, M. Machhout, L. Khriji, A. Baganne, and R. Tourki, "A modified aes based algorithm for image encryption," *International Journal of Computer Science and Engineering*, vol. 1, no. 1, pp. 70–75, 2007.

[33] D. S. A. Elminaam, H. M. Abdual-Kader, and M. M. Hadhoud, "Evaluating the performance of symmetric encryption algorithms." *IJ Network Security*, vol. 10, no. 3, pp. 216–222, 2010.

[34] J. W. Yoon and H. Kim, "An image encryption scheme with a pseudorandom permutation based on chaotic maps," *Communications in Nonlinear Science and Numerical Simulation*, vol. 15, no. 12, pp. 3998–4006, 2010.

[35] K. Gaj, E. Homsirikamol, and M. Rogawski, "Fair and comprehensive methodology for comparing hardware performance of fourteen round two sha-3 candidates using fpgas," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2010, pp. 264–278.

[36] K. Aoki, T. Ichikawa, M. Kanda, M. Matsui, S. Moriai, J. Nakajima, and T. Tokita, "Camellia: A 128-bit block cipher suitable for multiple platformsdesign andanalysis," in *International Workshop on Selected Areas in Cryptography*. Springer, 2000, pp. 39–56.

[37] S. Anis *et al.*, "Fpga implementation of parallel particle swarm optimization algorithm and compared with genetic algorithm," *International Journal of Advanced Computer Science & Applications*, vol. 1, no. 7, pp. 57–64.

[38] A. Webster and S. E. Tavares, "On the design of s-boxes," in *Conference on the Theory and Application of Cryptographic Techniques*. Springer, 1985, pp. 523–534.

[39] M. Abdelwahab *et al.*, “High performance fpga implementation of data encryption standard,” in *Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE), 2015 International Conference on*. IEEE, 2015, pp. 37–40.

Method for Game Development Driven by User-eXperience: A Study of Rework, Productivity and Complexity of Use

Mario González-Salazar*, and Hugo Mitre-Hernández†
Software Engineering Group,
Center for Research in Mathematics (CIMAT)
Av. Universidad 222, 98068
Zacatecas, Mexico

Carlos Lara-Alvarez‡
CONACYT Research Fellow
Center for Research in Mathematics (CIMAT)
Av. Universidad 222, 98068
Zacatecas, Mexico

***Abstract*—The growing capabilities and revenues of video game development are important factors for software companies. However, game development processes could be considered immature, specifically in the design phase. Ambiguous requirements in game design cause rework. User-eXperience (UX) is usually assessed at the end of the development process, causing difficulties to ensure the interactive experience between the game and users. To reduce these problems, this paper proposes a method for Game Development driven by User-eXperience (GameD-UX) that integrates a repository based on requirements engineering, a model for user experience management, and an adjusted agile process. Two experiments were conducted to study rework and productivity of video game development. Results of the first experiment revealed that GameD-UX causes less rework than conventional approaches, but it induces lower productivity. A tool for supporting the GameD-UX method was developed by considering the lessons learned. The second experiment showed that the software tool increases the productivity and reduces the complexity of use of GameD-UX.**

Keywords*—Rework; Productivity; Complexity of Use; Video Game Development

I. Introduction

Video games are important economically, they constitute the main entertainment industry, with continuous growth and billions of dollars in sales and revenues [1]. CEO of the Entertainment Software Association (ESA) point out that “Video games are the future; from education and business, to art and entertainment, our industry brings together the most innovative and creative minds to create the most engaging, immersive and breathtaking experiences we’ve ever seen...” [1]. However, ensuring the correct level of interactive experience between the game and the player is a challenge [2]; additionally, 65% of problems in game development are generated at pre-production stage and are related to unspecified or ambiguous requirements in game design [3], [4] causing rework and low productivity.

This article presents the *Game Development driven by User-eXperience* (GameD-UX) method that is composed by an improved Game Design Document (iGDD) [5] from requirements engineering perspective, a model for Game Experience Management (GEM) [6] from software architecture approaches, and an adapted agile method for game development. As shown in [6] the GEM model is able to improve the User eXperience (UX).

Besides defining the game requirements, the iGDD formalizes the game design by using software requirement principles. The main idea behind the GEM model is to transform the desired experience into game attributes. The quality attributes in conventional software products are security, usability, performance, among others; but in UX, these attributes can be seen as factors such as enjoyment, excitement, frustration, boredom, fear and more. Finally, the iGDD and GEM were included into a modified agile model for game development.

Two experiments were conducted to compare rework, productivity, and complexity of using GameD-UX and a conventional approach to game development. Results of the first experiment revealed that GameD-UX causes less rework, but it also induces lower productivity. The main difficulties found were attributed to failures in capturing and querying information from the iGDD and GEM – i.e. the low productivity was caused mainly by the complexity of using text documents; to overcome these limitations, a tool for supporting the GameD-UX method was developed. The second experiment confirmed that the GameD-UX supported by an appropriated tool produces better results in terms of rework, productivity, and complexity of use.

The rest of this article is organized as follows: section II presents the work related to the game repository, development and UX evaluation. Section III presents the method for Game Development driven by User-eXperience and its tool. Section IV explains the experiments. Section V presents and discusses the results. Finally, conclusions are presented in section VI.

II. Related Work

This section presents the related work in video game development (repository, UX evaluation, and development models), and how the proposed method alleviates the problems found in conventional approaches.

The game development process is composed by three stages: pre-production, production, and post-production [7]. The pre-production stage focuses mainly on creating the game concept and design. The production stage creates and validates the software; this stage also produces visual and auditory assets required by the game. Finally, the post-production stage distributes and maintains the game; it also manages the feedback coming from different sources – e.g., reviews.

A. Game Design Repository

Games are complex systems requiring significant effort in the first two development stages – pre-production and production. These complexities can increase the amount of rework and consequently, the cost of the game. The rework can be avoidable in most cases by detecting and correcting problems in early stages. A game design document can help to specify and structure the requirements of the game. The following sections have been proposed in the literature:

- **Overview.** Almost all authors suggest that a GDD should include a section that summarizes the key elements of the game to keep the eyes on the road [8]. Some authors even include a subsection of goals or objectives of the game [8], [9], [10], [11], [12], [13], [14].
- **Mechanics.** The term *mechanics* is used to describe game elements – e.g., player characters – and interaction rules – e.g., a player-enemy interaction. Mechanics include characters or assets list [8], [9], [10], [11], [12], [13], [14].
- **Dynamics.** Several proposals have common sections that contains intended interactions with the player such as interfaces, levels or challenges [8], [9], [10], [11], [12], [13], [14].
- **Aesthetics.** It is what the player perceives by his visual and auditory senses. Most authors only cover the visual aspects in a document called the art bible. Baldwin [12] suggests a GDD template that abbreviates an art bible. Auditory assets can also be included in this section [8], [9], [14].
- **Experience.** Creating enjoyable player experience is fundamental for the game success [2]. Player experiences are enriched by mechanics, dynamics and aesthetics of the game. Playability can be used to link game design to player experience [15]. Therefore, defining the expectations of player experiences may lead to the improvement of the game and to the establishment of a base line to test the experiences in production.
- **Assumptions and Constraints.** Some authors include technical limitations in the technical bible [10], [12], [14].

An effort to integrate the previous sections into a single repository, the authors of this paper have proposed the improved Game Design Document (iGDD) [5]; iGDD sections are related to the Software Requirement Specification (SRS) characteristics as described in Table I. The method proposed in this paper also uses the iGDD as repository.

TABLE I. DESCRIPTION AND CHARACTERISTICS OF SECTIONS OF THE iGDD

iGDD Section	Description	SRS Characteristics
Overview	Describes briefly the most important aspects of the game.	Relations with other documents, and common language for better understanding
Mechanics	Describes the elements of the game.	Organization of game requirements (objects organization).
Dynamics	Describes how the elements of the game will take action in the game.	Organization of game requirements. Relation of complexity with gamer profile.
Aesthetics	Describes what the player perceives directly through their sense, like what he sees and hears.	It is not related to the SRS.
Experience	Highlights important aspects of the game and what you hope to achieve from these aspects.	Decision-making based on trade-offs of game parts. Quality attributes on video games.
Assumptions and constraints	Narrates the aspects of the design assumptions and limitations of the game, either technical or business.	Knowledge of game parts for reviews. Limitations or boundaries of video game

B. Game User eXperience Evaluation

Guaranteeing an enjoyable User eXperience (UX) is critical for game companies. There are some related works seeking to solve the problem:

- **Core Elements of the Gaming Experience (CEGE)**. Calvillo et al. [16] suggest that core elements to ensure UX are: puppetry (control, ownership, and facilitators) and video game (game-play and environment). They also propose a questionnaire for evaluating these elements.
- **Game Experience Questionnaire (GEQ)**. Engl and Nacke [17] consider that immersion, flow, competence, tension, challenge, positive and negative affect are UX evaluation factors. They also propose a comparative evaluation instrument.
- **Heuristics.** Hochleitner et al. [18] propose a framework of heuristics (design guidelines for aesthetics and mechanics in a game genre) categorized in game play/story, and virtual interface to asses UX.

Although heuristics are part of the game design, they are considered general guidelines that belong to a game genre; similarly, the GEQ instrument does not ensure the UX because it can be only used after the game is finished. Conversely, integrating CEGE components for designing the game could avoid unpleasant experiences. In a previous work, the CEGE was compared to the *Game Experience Management Model*; games developed using the GEM, improves the UX [6].

The GEM is based on software architecture because of its advances in software systems design. The interaction experience between game and player can be interpreted as a set of quality attributes in software engineering [2]. In traditional software systems, quality attributes include: security, usability, performance, etc. In video games these attributes could be considered as factors of User eXperience (UX) as: enjoyment, excitement, attention; these attributes are closely related to emotions and cognitions of the player.

The quality attribute approaches in software architecture design can be categorized into [19]:

- **Quality Attribute Requirement Focused (QARF).** These approaches perceive Quality Attribute requirements as the main focus in the software architecture design phase, and consider each design decision based on its implications on the prioritized quality attributes [20];

- **Quality Attribute Scenario Focused (QASF).** These approaches map architectural quality goals into concrete scenarios to characterize stakeholders concerns throughout the software architecture design phase [20], [21];
- **Influencing Factor-Focused (IFF).** These approaches focus on the inter-dependencies among factors and constraints that would affect the choice of design decisions [21].

Software architecture design is an area that can bring the solution idea for UX management in pre-production stage. The QARF and QASF approaches are suitable for game design due to its aspects of quality attributes requirements, prioritized QA, and the scenarios of design. Into these categories we can find the Quality Attribute Workshop (QAW) and the Attribute Driven Design (ADD) methods. The QAW [20] is a facilitated, early intervention method used to generate, prioritize, and refine quality attribute scenarios before the software architecture is completed. The ADD defines software architecture by basing the design process on the quality attributes that the software must fulfill [21].

Initially, the GEM model [6] defines a high-level game design that associates game goals of the iGDD with the desired experience; the experience is described in design drivers, detailed in guidelines and verified by test cases. A design driver is a high-level property that the game should have in order to generate the intended experience in the player; i.e., the user experience metrics – emotions as fear, happiness, angry, etc. In each iteration of game development, game elements are created and checked to confirm that the game is achieving the goals. A game design guideline is a description of how game elements need to be created in order to achieve the intended experience established in the game design drivers. Finally, the test cases evaluate guidelines in terms of fulfillment of their goals; it could include a questionnaire or an emotional evaluation model and its relation to parts of the game. There must be at least one test case per guideline. In sum, GEM is able to design and manage the expected UX in the proposed method.

C. Video game development models

The video game development is a form of software development that adds additional requirements, – e.g., artistic aspects; hence, many of the management tools and standards from the software industry can be useful for game development. Game projects are usually more complicated than software projects because they involve a multidisciplinary team and they usually have more uncertainty around project goals. Software development models – e.g. waterfall, iterative, or extreme can be used for developing video games [7]. In general, the waterfall model is considered inadequate because it is highly structured and it cannot be adapted to changes in the requirements; therefore, more flexible models are needed [22], [2], [23].

Agile methodologies – i.e. Scrum [24] or eXtreme Programming [25] – are better suited for the challenges of game development [26], [27]; they have been adapted to game development by using other tools as complements: user stories [26], game design documentation [28], or workshops for strengthening the interaction between clients and developers [29].

The adjustment of software development to a specific context is well studied in software engineering through patterns. Patterns [30] are used to solve a generic problem: given a narrative and context of the problem to be solved, they propose a solution. They can be used for formalizing the knowledge about the development process. In [31] the authors propose the Software Development Project Pattern (sdPP) framework. For testing this approach [31], generates four instances of the sdPP with agile development models; one of these instances Scrum sdPP is suitable for game agile development because it allows to follow an iterative process without sacrificing creativity. The resulting workflow and productflow can guide game developers between the activities and their corresponding input and output products.

An sdPP instance of Scrum was adapted for agile game development. In this modified pattern instance, the iGDD and GEM were integrated. The main activities in relation with iGDD and GEM are described in the proposed method.

III. Proposed Approach

This section first describes the GameD-UX activities and how they are related to the iGDD and the GEM, then it describes the software tool improvements based on the lessons learned of the initial experiment.

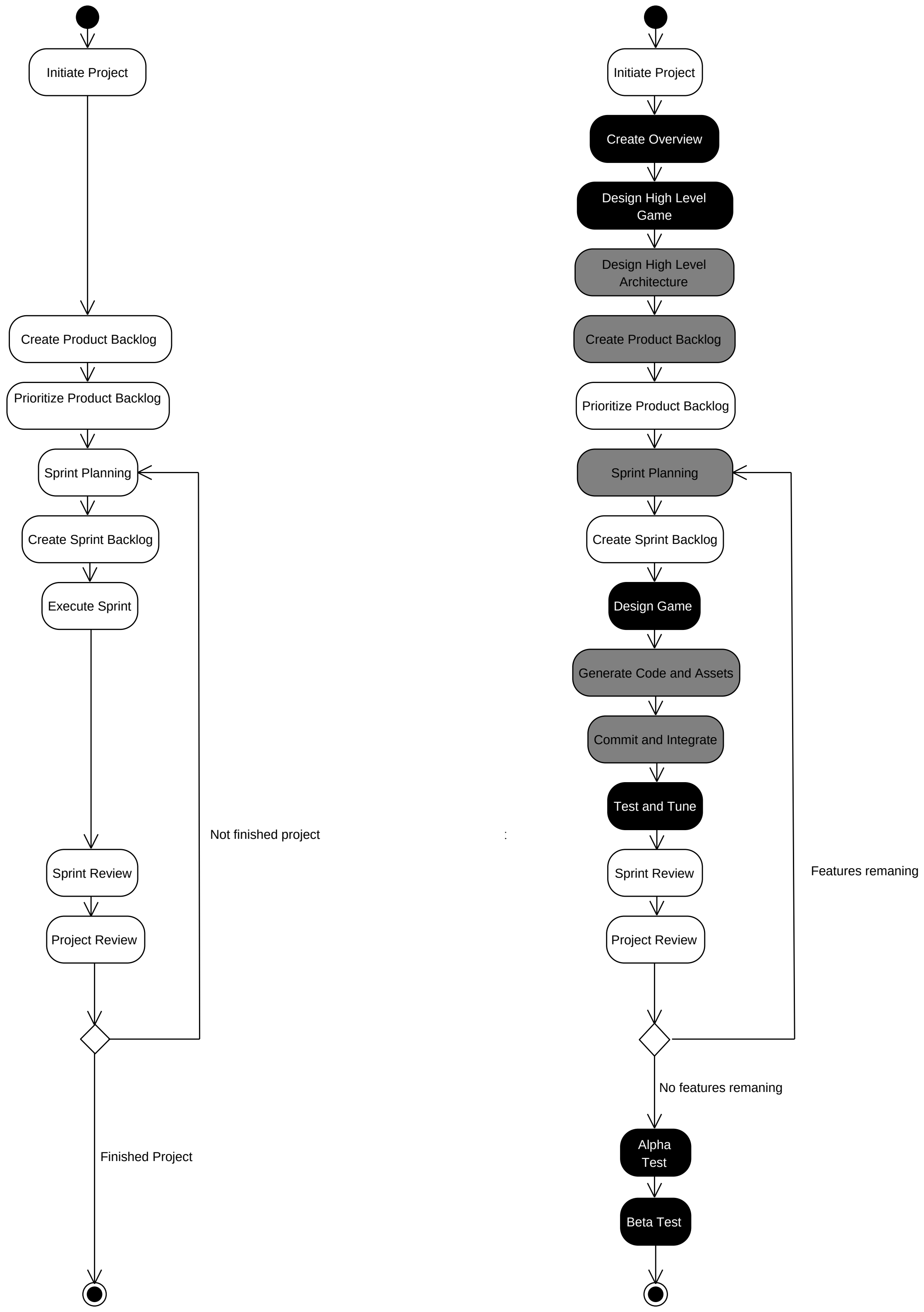
Game Development driven by User-eXperience (GameD-UX) is a method to design and develop video games from the required user experience. It uses two components: (i) the improved Gamed Design Document (iGDD) for the repository of all structured game elements, and (ii) the Game Experience Management (GEM) model to capture and manage the required UX along game development.

A. Method for Game Development driven User-eXperience

GameD-UX is composed of a repository containing game design (iGDD); a model to design, track and manage user experience (GEM); and an adapted Scrum method for game development, with the aim to design and develop video games based on user experience. Scrum was selected because is a flexible framework that can be adapted to other methods or tools – e.g., user stories, Kanban board –, its cycle life is iterative, incremental, and evolutionary. It does not sacrifice creativity, and it is well documented [26].

The GameD-UX activities are based on the general Scrum activities. Fig. 1 illustrates these activities and their relationship to iGDD and GEM. The following paragraphs describe these activities:

1) **Initiate the project.** A game development project may have different sources: an original idea for a game, or an opportunity found in a specific market. Once a game idea from some source initiates a game development project, the first activity is to assign resources and to transform the game idea into the game concept.



(a)

(b)

Fig. 1. Comparison of the general activities between: (a)Scrum, and (b)the proposed approach. The proposed approach adds (black) and modifies (gray) activities.

2) **Create overview** *(product: overview of the iGDD)* The overview describes the game in a brief abstract, identifies the main objectives of the game, the genre of the game, asks questions e.g., why the game is

worth doing, defines which type of players would like to play the game, and what will be the main activities that the player will be doing while playing the game.

3) **Design high level game** *(product: overview of the iGDD, drivers and guidelines of the GEM).* The high level game defines some main features of the game: the game modalities (single player, multiplayer, online, arcade mode, history mode, among others), the platform or platforms on which the game is intended to run, the game theme (medieval, futuristic, western, among others), the game story and an initial scope of the levels, size and time that the game may require. Based on the overview information, it defines the high level properties (drivers) that the game should have in order to bring the desired UX. For each driver, it defines one or more guidelines on how to create specific game element(s) in order to fulfill the driver goal. Team members must approved guidelines.
4) **Design high level game architecture** *(product: assumptions and constraints, mechanics, dynamics of the iGDD and guidelines of the GEM).* The high level architecture reviews the technical settings to modify the assumptions and constraints. Technical settings include: the standards, conventions, technology, resources and architecture selected for the game. This activity creates a high level version of the game elements related to the guidelines.
5) **Create product backlog** *(product: overview of the iGDD).* The main features in the game listed in the overview are used to create the requirements in the product backlog.
6) **Prioritize product backlog** *(no iGDD or GEM section associated).* The team prioritize the product backlog requirements based on the value that each requirement give to the game.
7) **Organize product backlog** *(no iGDD or GEM section associated).* The product backlog lists everything that might be needed in the game, the resulting list has the requirements to be implemented in the project.
8) **Sprint planning** *(no iGDD or GEM section associated).* In this activity, the requirements with the higher priority from the backlog are selected and planned.
9) **Create sprint backlog** *(no iGDD or GEM section associated).* In this activity, each task derived from the chosen requirements is estimated and assigned to the team members as long as there is time left in the sprint.
10) **Design Game** *(products: mechanics, dynamics of the iGDD and guidelines of the GEM).* This activity designs each game element needed to fulfill the requirements to be developed on the sprint and verifies that the game elements follow the guidelines associated to them (if there is any). It also validates that designed game elements correspond to guidelines. Finally, it creates test cases to validate guidelines (if needed).
11) **Generate code and asset** *(products: mechanics, dynamics of the iGDD and guidelines of the GEM).* This activity creates game elements based on the game design and their corresponding guidelines. These elements include code and assets – e.g., music or animations.
12) **Commit and integrate** *(products: guidelines of the GEM).* This activity validates that the developed game elements follow the guidelines or gives a valid justification of why they could not follow them. It also integrates game elements in a version suitable for release.
13) **Test and tune** *(products: guidelines and test cases of the GEM).* This activity tests the resulting product of the sprint in order to verify the quality e.g., fulfill the guidelines. Small adjustments can be made to polish the game, but radical changes should be placed in the product backlog to consider them in the next sprint. The result of this activity will be a potentially shippable product.
14) **Sprint review** *(products: GDD all and GEM all).* The retrospective presents the results of the sprint: reviews of the product, process, tools, people, and any other relevant aspect of the project. The feedback given by members of the team and other stakeholders is evaluated.
15) **Project review** *(products: GDD all and GEM all).* The information of previews sprints is used to evaluate the project, if needed the team adjusts the project duration; modifies, eliminates or adds requirements in the product backlog. While there are pending requirements in the backlog go to activity 7.
16) **Do alpha test** *(products: test cases of the GEM).* Alpha test finds and removes bugs and verifies that game elements fulfill quality criteria [26], [32], [11].
17) **Do Beta test** *(products: test cases of the GEM).* This test evaluates UX.

B. Software tool to support GameD-UX

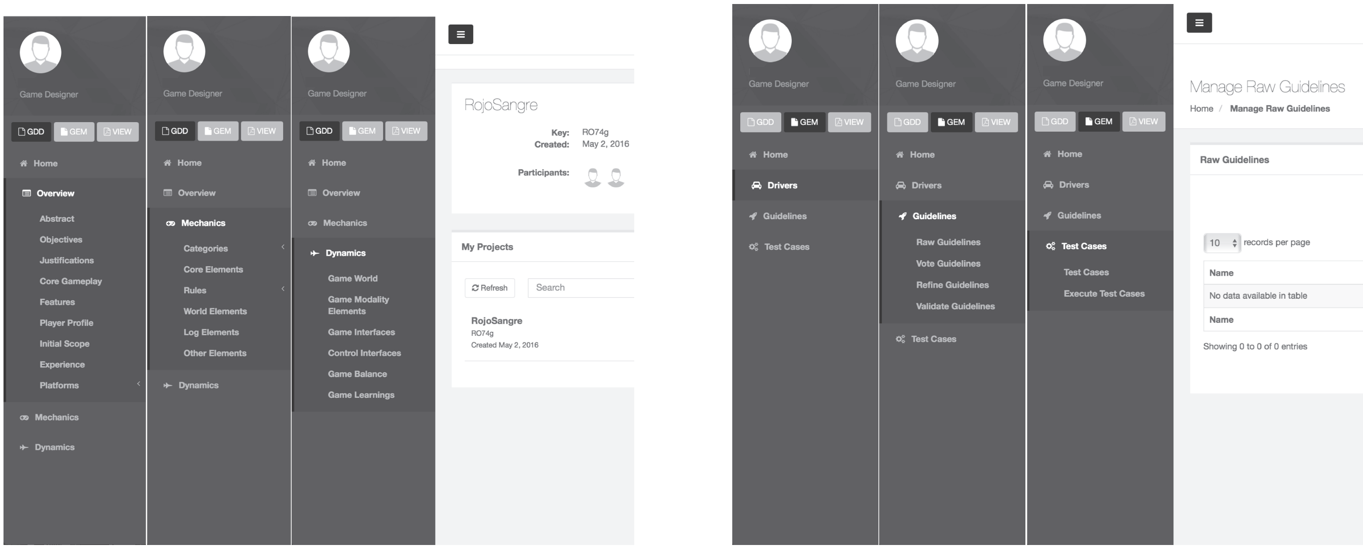
GameD-UX can help to improve the experience of the player [6] and reduce the rework [33] of the game development team. Nevertheless, in opinion of developers after the execution of the first experiment, the complexity of using iGDD and GEM together provokes low productivity. For this reason, we designed a software tool aiming to reduce the complexity of use of GameD-UX; hence, the requirements presented in Table II were defined.

The tool to support GameD-UX has two menus: the iGDD (Fig. 2a) and the GEM (Fig. 2b). The iGDD menu can create, modify or disable game categories and elements. The game designer can change the status of a category or an element. The tool enforce to follow the structure of the iGDD – e.g., if the user wants to create a ninja, it is necessary to create the enemies category.

Analogously, the GEM menus enforce to follow the GEM structure. It is necessary to have a goal in the iGDD overview to associate a driver, a driver to create a guideline, and so on.

IV. Material and Methods

Two experiments are presented in this section following the suggestions of Wohlin et al. [34]. The purpose of the first experiment is to evaluate GameD-UX in terms of rework and productivity. The second experiment is intended to evaluate the productivity and complexity of using the GameD-UX tool.



RojoSangre

Key: RO74g
Created: May 2, 2016
Participants:

RojoSangre
RO74g
Created May 2, 2016

(a)

Manage Raw Guidelines

Home / Manage Raw Guidelines

Raw Guidelines

10 records per page

Name
No data available in table
Name

Showing 0 to 0 of 0 entries

(b)

Fig. 2. Software Tool: (a) menus that support the iGDD. (b) menus that support the GEM.

TABLE II. SOFTWARE TOOL REQUIREMENTS

N	Description
R_1	The tool can have information on the status of each element created in the iGDD, to know if an element is in progress (which means it can't be implemented yet) or it is already finished (which means ready to implement).
R_2	The tool can keep track of the game elements in the iGDD with the drivers and guidelines in the GEM and show with which guidelines and drivers the game elements are linked.
R_3	The tool can filter of all the information of the game elements in the iGDD, to let the team know which elements are finish (ready to implement) and which are in progress. It can also help to filter the information by role (it let know which game elements are related to the art roles or software engineering roles).
R_4	The iGDD in the GameD-UX method is a repository with a precise structure and relations, by ensuring with a tool that the game designers keep this structure and relations. Requirements can be easily linked to the game element by the method taxonomy. If the requirement involves the development of a core game element, or a challenge, these key categories can be present in the requirement description as a tag with a color associated to them.
R_5	The tool can ensure that the game elements are created correctly following the structure and inter-dependencies pre-determined. For instance, if the game designer wants to create a challenge, but there are no core game elements, the challenge will be empty until the core game elements are created and can be integrated in the challenge. Same way, if the game designer want to create a core game element but there is no category to which this must belong the element cannot be created. In sum, the constraints of iGDD elements and its relations are automated in the tool.
R_6	The tool can notify with which guidelines (GEM elements) the game elements are related and what the guideline says about the game element to be developed. This makes easy to confirm if the game element follows the guideline(s).

A. Context

Students who successfully completed the *video game development course* in the software engineering master degree program, in the Center for Research in Mathematics (CIMAT) were eligible to participate in the study. Besides the fundamental concepts of video game development, this course also focuses in Unity® as game engine and C# as programming language.

B. Participants

Eighteen junior software engineers carried out this empirical study; these engineers (hereafter, participants) had experience in Scrum. Three groups were considered:

Group A. Three development teams of two participants that use GameD-UX.
Group B. Three development teams that use the conventional approach composed by: Taylors GDD [8], and the agile game development with Scrum [26].
Group C. Three development teams that uses the GameD-UX tool.

Each team (composed of two participants) developed a single video game. The game overview can be summed as: a tower defense game for teaching basic multiplication operations to children in elementary school. The goal of a tower defense game (a special case of strategy video games) is to stop enemies from reaching a specific point on a map; for this, the player can build towers to kill enemies.

C. Metrics

a) Rework: is defined as any additional effort required for finding and fixing problems after documents and code are formally signed-off as part of configuration management [35]. For measuring the rework, any artifact put to test for the first time starts to register rework time after the test is done. To compare different products, rework effort is sometimes normalized by being calculated as a percentage of development effort [35].

b) Productivity: is the amount of requirements that a team can complete in an hour.

c) Complexity of use: of GameD-UX tool is evaluated with a post-mortem survey applied to participants (teams A and C) after finishing the project. The survey evaluates the complexity of using GameD-UX with the software tool, it contains the following assertions evaluated in a likert scale (1 strongly disagree – 7 strongly agree):

Assertion 1 (A_1): The tool or text documents easily allow to associate the guidelines (GEM) with the game elements (iGDD) to be developed.

Assertion 2 (A_2): The tool or text documents makes easy to identify which game elements (iGDD) are in progress and which are finished.

Assertion 3 (A_3): The tool or text documents facilitates the validation of game elements (iGDD) with their corresponding guidelines (GEM).

An overview of the two experiments (A, B) and their relationships to the study groups, metrics, and lesson learned are shown in Fig. 3.

D. Experiments

The experiment A was conducted to compare rework and productivity of GameD-UX and a conventional approach to game development. For this aim, projects of groups A and B are compared in terms of rework and productivity.

The experiment B evaluates the software tool developed for supporting GameD-UX in terms of productivity and the complexity of use. For this aim, projects of groups A and C are compared in terms of productivity and the post-mortem survey.

V. Results

Experiment A

As shown in Fig. 4, the normalized mean of rework for group A was 2.73%, while for group B was 11.50%. A Wilcoxon signed-rank test showed that the GameD-UX induces significantly less rework than the induced in the group B ($p < 0.01$). This proves that GameD-UX generates less rework than the conventional approach.

Concerning productivity, the mean of requirements finished in Group A was 11, the mean in Group B was 11.66; the productivity mean in Groups A was 0.22 requirements/hour and in Group B was 0.29. A Wilcoxon signed-rank test showed that GameD-UX (group A) induces significantly less productivity than the productivity induced in the group B that used the conventional approach ($p < 0.01$). This means that GameD-UX without supporting tool has the disadvantage of low productivity in comparison with the conventional approach. To illustrate this disadvantage of our method, Fig. 5 shows a boxplot of productivity distribution in both groups.

After the groups delivered their developments, a post-mortem evaluation was performed to evaluate the good and bad experiences of using the main elements of GameD-UX. The post-mortem analysis of the game developers experiences (Group A) brings us potential evidences to explain the poor productivity in our method. The resumed lessons learned of the six participants that use the iGDD and GEM are:

- It is unclear when to modify the iGDD content.
- There exists a missing link between requirements and GEM guidelines.
- It is hard to recognize elements to be developed by role of the team; e.g., artist, programmer, designer.
- It is hard to visually associate requirements to game elements of the iGDD.
- It is hard to recognize the correct sequence of sections to be developed.
- Developers are usually overloaded with manual validation of GEM elements.

TABLE III. Productivity (Hours) for Groups A and C

	Requirement Media						
	R_1	R_2	R_3	R_4	R_5	R_6	Total
Group A	21.5	32	12	17	31	18	131.5
Group C	20	31	10	16	28	11	116.0

Experiment B

Results of experiment A show that manual work (using text documents) to manage GEM, iGDD, and the links between them generates low productivity for video game development. To overcome this issue, a tool (described in section 3.2) was developed following the lessons learned.

In general, the productivity of group C (total median of 116.0 hours) is better than the productivity of group A (total median of 131.5 hours). Table III shows the results for each requirement.

The complexity evaluation was measured by applying an independent sample t-test to examine if there was a significant difference among the means of assertions answers. The boxplot shows the median results for the survey (Fig. 6). A statistical difference was observed for A_1 ($p < 0.01$) and A_2 ($p < 0.01$). It means that the supporting tool facilitates the use of GameD-UX because it easily associates guidelines with game elements and it makes easy to identify which game elements are in progress and which are finished.

Although the scale for A_3 was higher when using the supporting tool (Mean=5.56) compared to the GameD-UX (Mean=4.63), there was not a significant difference.

VI. Conclusions

This paper presents the GameD-UX method for video game development based on UX. It is composed of a repository of game elements (iGDD), a model to design, track and manage user experience (GEM), and an adapted Scrum method for game development.

The GameD-UX method induces less rework than a conventional approach used to develop video games (Taylors GDD and the agile game development with Scrum). Sections of the iGDD and their relation to the Software Requirement Specification (SRS) characteristics are key factors that improve the conventional repository.

The lessons learned from initial video game developments using GameD-UX – e.g., low productivity – were improved with the supporting tool. The requirements developed by teams that uses the GameD-UX consumes 13.33% less time when using the tool.

According to the survey applied to game developers, the complexity of use of GameD-UX was reduced with the tool. Specifically, the GameD-UX tool: (i) allows to associate the

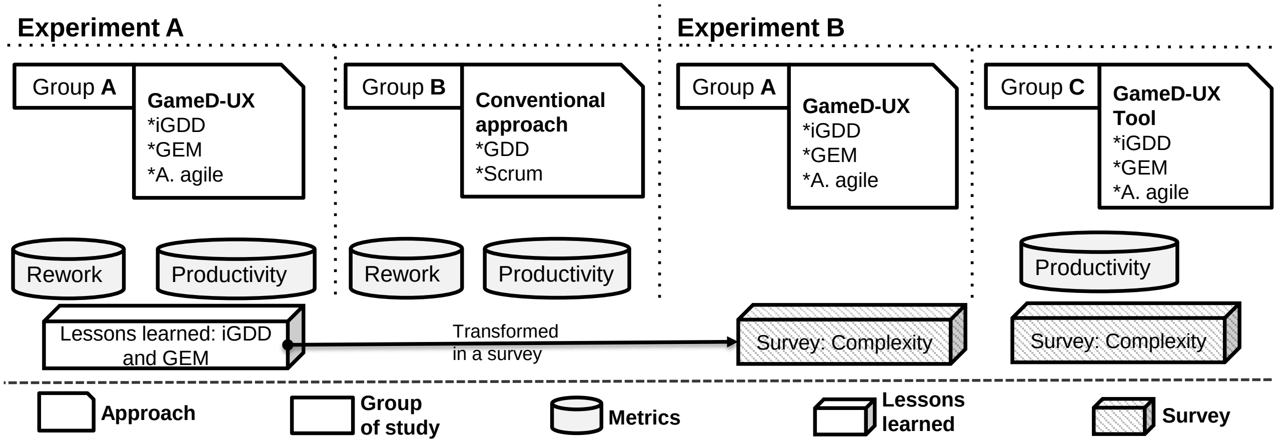


Fig. 3. Groups, metrics and instruments for experiments A and B.

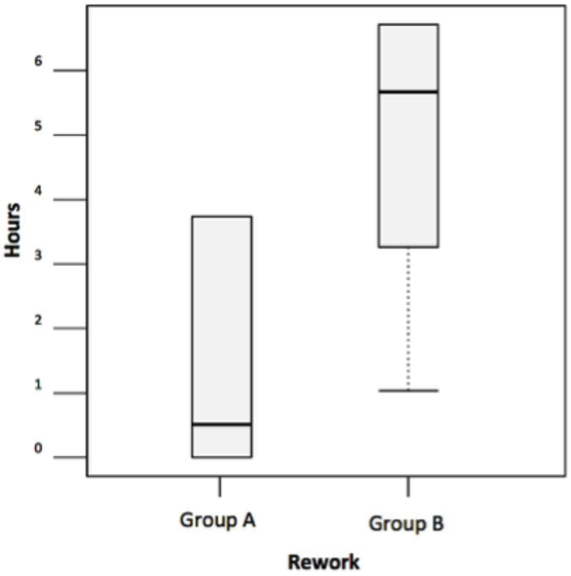


Fig. 4. Comparison of rework for groups A and B.

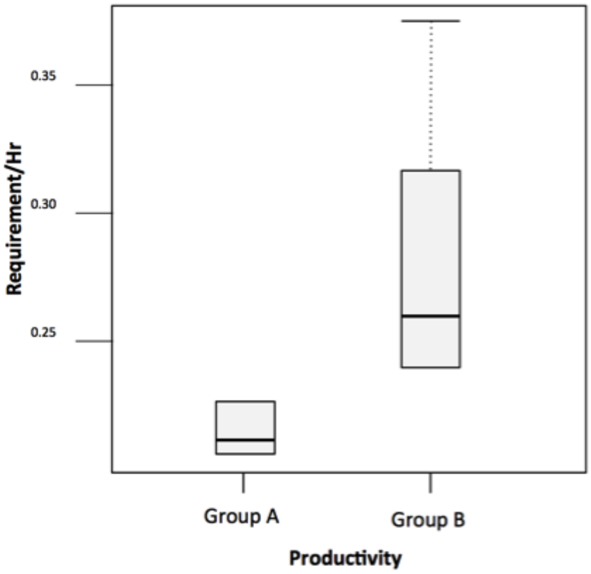


Fig. 5. Comparison of productivity for groups A and B.

guidelines (GEM) with the game elements (iGDD), and (ii) it facilitates the identification of game elements according to their status (finished or unfinished). The tool was designed to track game elements and their association to guidelines and drivers. But we believe that a better performance can be obtained by improving the tool – e.g., including real-time notifications of status changes to developers and reviewers.

In further works, we will extend the GEM to involve player-centric practices – e.g., players can help to define their profile. A more accurate understanding of the potential players, will originate more useful game design drivers to create better an experience for these players.

In affective and cognitive computing, we want to investigate the human behavior with the video game elements in order to achieve the desired emotion or cognition, and integrate it in the GEM the comparative results of diferent versions of mechanics and asthetics.

References

[1] E. S. Association *et al.*, “Essential facts about the computer and video game industry: 2010 sales, demographic and usage data 4 (2010),” *Washington, DC. Disponível em: http://www. theesa. com/wp-content/uploads/2014/10/ESA_EF_2014. pdf. Acesso em mai*, 2016.

[2] J. Schell, *The Art of Game Design: A book of lenses*. CRC Press, 2014.

[3] F. Petrillo, M. Pimenta, F. Trindade, and C. Dietrich, “What went wrong? a survey of problems in game development,” *Computers in Entertainment (CIE)*, vol. 7, no. 1, p. 13, 2009.

[4] ——, “Houston, we have a problem...: a survey of actual problems in computer games development,” in *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 2008, pp. 707–711.

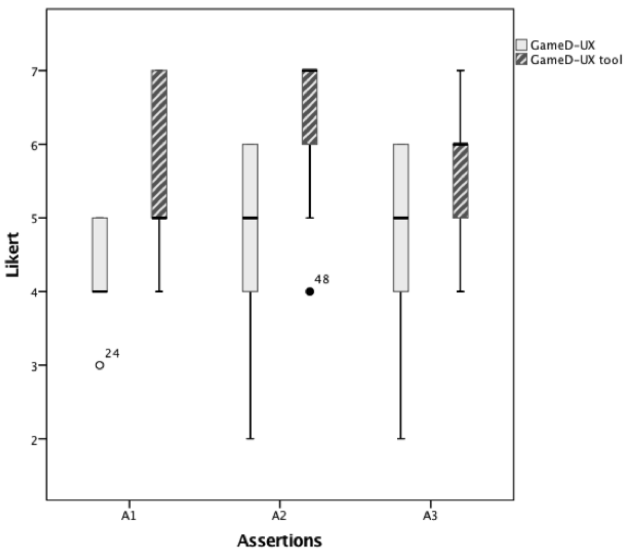


Fig. 6. Comparison of complexity of use GameD-UX between groups A (GameD-UX) and C (GameD-UX tool)

[5] M. Gonzalez-Salazar, H. A. Mitre, C. L. Olalde, and J. L. G. Sánchez, "Proposal of game design document from software engineering requirements perspective," in *Computer Games (CGAMES), 2012 17th International Conference on*. IEEE, 2012, pp. 81–85.

[6] H. Mitre-Hernandez, C. Lara-Alvarez, M. Gonzalez-Salazar, J. Mejia-Miranda, and D. Martin, "User experience management from early stages of computer game development," *International Journal of Software Engineering and Knowledge Engineering*, vol. 26, no. 08, pp. 1203–1220, 2016.

[7] E. Bethke, *Game development and production*. Wordware Publishing, Inc., 2003.

[8] C. Taylor. (1999) MS Windows NT design document. [Online]. Available: www.designersnotebook.com/ctaylordesign.zip

[9] S. Rogers, *Level Up! The guide to great video game design*. John Wiley & Sons, 2014.

[10] K. Oxland, *Gameplay and design*. Pearson Education, 2004.

[11] A. Rollings and E. Adams, *Andrew Rollings and Ernest Adams on game design*. New Riders, 2003.

[12] M. Baldwin. (2005) MS Windows NT game design document outline. [Online]. Available: http://ccasummer2014.tumblr.com/post/91982395367/baldwin-game-design-document-template-doc-file

[13] B. Bates, *Game Design [Paperback]*. Premier Press; 2nd Revised edition edition, 2004.

[14] R. Rouse III, *Game design: Theory and practice*. Jones & Bartlett Learning, 2010.

[15] L. Nacke, A. Drachen, K. Kuikkaniemi, J. Niesenhaus, H. J. Korhonen, W. M. Hoogen, K. Poels, W. A. IJsselsteijn, and Y. A. De Kort, "Playability and player experience research," in *Proceedings of DiGRA 2009: Breaking New Ground: Innovation in Games, Play, Practice and Theory*. DiGRA, 2009.

[16] E. H. Calvillo-Gámez, P. Cairns, and A. L. Cox, "Assessing the core elements of the gaming experience," in *Game User Experience Evaluation*. Springer, 2015, pp. 37–62.

[17] S. Engl and L. E. Nacke, "Contextual influences on mobile player experience–a game user experience model," *Entertainment Computing*, vol. 4, no. 1, pp. 83–91, 2013.

[18] C. Hochleitner, W. Hochleitner, C. Graf, and M. Tscheligi, "A heuristic framework for evaluating user experience in games," in *Game User Experience Evaluation*. Springer, 2015, pp. 187–206.

[19] H. P. Breivold, I. Crnkovic, and M. Larsson, "A systematic review of software architecture evolution research," *Information and Software Technology*, vol. 54, no. 1, pp. 16–40, 2012.

[20] M. R. Barbacci, R. J. Ellison, A. Lattanze, J. Stafford, C. B. Weinstock, and W. Wood, "Quality attribute workshops," 2002.

[21] R. L. Nord, W. G. Wood, and P. C. Clements, "Integrating the quality attribute workshop (qaw) and the attribute-driven design (add) method," DTIC Document, Tech. Rep., 2004.

[22] R. Hunicke, M. LeBlanc, and R. Zubek, "Mda: A formal approach to game design and game research," in *Proceedings of the AAAI Workshop on Challenges in Game AI*, vol. 4, no. 1, 2004.

[23] B. W. Boehm, "A spiral model of software development and enhancement," *Computer*, vol. 21, no. 5, pp. 61–72, 1988.

[24] K. Schwaber and M. Beedle, *Agile software development with Scrum*. Prentice Hall Upper Saddle River, 2002, vol. 1.

[25] K. Beck, *Extreme programming explained: embrace change*. addison-wesley professional, 2000.

[26] C. Keith, *Agile game development with Scrum*. Pearson Education, 2010.

[27] J. Kasurinen, R. Laine, and K. Smolander, "How applicable is iso/iec 29110 in game software development?" in *International Conference on Product Focused Software Process Improvement*. Springer, 2013, pp. 5–19.

[28] A. Godoy and E. F. Barbosa, "Game-scrum: An approach to agile game development," *Proceedings of SBGames*, pp. 292–295, 2010.

[29] R. Kortmann and C. Harteveld, "Agile game development: lessons learned from software engineering," in *Learn to Game, Game to Learn; the 40th Conference ISAGA*, 2009.

[30] C. Alexander, *The timeless way of building*. New York: Oxford University Press, 1979, vol. 1.

[31] D. Martín, J. G. Guzmán, J. Urbano, and J. Llorens, "Patterns as objects to manage knowledge in software development organizations," *Knowledge Management Research & Practice*, vol. 10, no. 3, pp. 252–274, 2012.

[32] I. van de Weerd, S. de Weerd, and S. Brinkkemper, "Developing a reference method for game production by method comparison," in *Situational method engineering: Fundamentals and experiences*. Springer, 2007, pp. 313–327.

[33] H. A. Mitre-Hernández, C. Lara-Alvarez, M. González-Salazar, and D. Martín, "Decreasing rework in video games development from a software engineering perspective," in *Trends and Applications in Software Engineering*. Springer, 2016, pp. 295–304.

[34] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[35] S. Pfleeger and B. Kitchenham, "Software quality: The elusive target," *IEEE Software*, pp. 12–21, 1996.

Face Recognition using SIFT Key with Optimal Features Selection Model

Taqdir
Assistant Professor, Computer Science and Engineering Department,
GNDU, Regional Campus, Gurdaspur-143521, Punjab, India

Renu Dhir
Associate Professor and Head, Computer Science and Engineering Department,
NIT, Jalandhar –144001 Punjab, India

***Abstract*—Facial expression is complex in nature due to legion of variations present. These variations are identified and recorded using feature extraction mechanisms. The researchers have worked towards it and created classifiers for identifying face expression. The classifiers involve Principal component analysis (PCA), Local Polynomial approximation (LPA), Linear binary pattern (LBP), Discrete wavelet transformation (DWT) etc. The proposed work deals with the new classifier using SIFT key with genetic algorithm to identify distinct facial expression. Optimal features of existing algorithms are used within the proposed work. Also comparison of existing techniques such as LBP, PCA and DWT is presented with SIFT key with genetic algorithm. The results show that proposed classifier gives better result in terms of recognition rate.**

Keywords—Feature Extraction; Classifier; PCA; LPA; LBP; DWT; SIFT key; Genetic algorithm

I. Introduction

Facial recognition system is a biomedical mechanism of identifying various expressions.[1] Facial recognition system is commonly used in security applications but also used heavily in other applications. Facial recognition system involves number of techniques. These techniques are primarily associated with feature extraction. Human face is a house of distinct expression which varies with time continually. Hence efficient classifier is required which generate number of optimal features as quantity to represent entire facial expression residing on human face. Optimal feature selection is difficult with single classifier hence properties of multiple classifiers are collaborated together to achieve optimal classifier.

[2]In this the non-domination based optimization technique has been introduced that recognize the known and unknown faces with a semi-supervised classifier that are based on the different scenarios. The identification is based on proper training sets with actual face images that provide reliable results. In this different datasets like Yale face database, ORL database has been utilized for experimenting and obtaining the robust results. The results are based on TP, FP and TN classification evaluation. It recognizes the faces by computing LNS of training sets that are considered using the grouping.

FACE RECOGINITION: When features are selected from an image than these feature uses to recognize the faces. In an automated face recognition system, huge diversity is found due to facial appearance during recognition of faces. So many systems are under development now a day that recognises the face based on the appearance. In this system the face image is represented in Eigen faces which consist of vectors of intensities.

There are following tasks which are associated with face recognition system:

- Verification
- Identification
- Watch List

VERIFICATION: The first task of face recognition system is associated with the access applications. The access applications are the one which provide user interaction towards the recognition system. The verification is the process to check the identification of the person that is being access the system using an application. The verification of the persons can be done using two groups:

1) CLIENTS: They are the persons that have access to the system with identity. The percentage rates at which clients are to be rejected are known as False Rejection Rate (FRR).

2) IMPOSTERS: They are the peoples which uses false identity to gain access to the system. With term false identity we mean the identities that are to belonging to the system but known. The imposter gaining access known as False Acceptance Rate (FAR).

IDENTIFICATION: This is done in surveillance application where applications do not need user interaction. It is based on the assumptions that all faces in the Image are known faces. The correct identification percentage can be measured as Correct Identification Rate (CIR) and false identification can be measured using False Identification Rate (FIR).

WATCH LIST: It is the generalized form of identification task in which we will include unknown persons also. To describe the sensitivity of the watch test it includes FAR and FRR along with the identification test reported in CIR and FIR. It describes how often an unknown person tries to access the system.

The proposed work has following finding associated with it.

- Determining optimal set of features from different training images.
- Comparing Accuracy of different classifiers on training images.
- Comparing different techniques for optimum feature extraction with distinct classifier.
- Comparing performance of different approaches in terms of recognition rate.

The proposed work begins by giving the introduction of classifiers used for face recognition. Next section describes proposed classifier with genetic algorithm. The next section gives the result associated with proposed system. Last section describes conclusion and future work.

II. Face Recognition using Principal Component Analysis

Face has distinct variations associated with it. Analysing facial expression at distinct time interval is a challenging task. One of the techniques used to analyse facial recognition is with the help of Principal Component Analysis.[3]PCA is a statistical procedure that uses orthogonal transformation to convert possibly correlated values into set of non-correlated linear values known as principal components. It utilizes the set of eigenvalues that are builds from the set of training data sets. From these eigenvalues the training face images have been calculated which are arranged for finding the most variance in image. After this the Euclidean distance from the input face has been calculated for each eigenvalues. This can classified the image into parts based on Euclidean distance. The weighted sum of eigenfaces represented by text face images projected on to the space expanded by eigenfaces. The faces can be identified by these weights.

The following is way through which the correlated values has been calculated:

$$\mu = \frac{1}{m}\sum_{n=1}^{m} x_n$$

$$C = \frac{1}{m}\sum_{n=1}^{m} (x_n - \mu)(x_n - \mu)^T$$

This approach is useful and found application in [4] where it is used to detect emotions. Health of human being is greatly influenced by emotion. Five kinds of emotions are detected in this study. Number of extracted features from the ECG study is reduced by the use of PCA. Also another study using this PCA technique is given in[5] showing emotion detection for movement assisted in wheelchair.

Face Recognition using LBP and LPA

Image is generally represented either as 2D or 3D objects. 2D objects representation is relatively less clear as compared to 3D objects. Images represented as 3D objects can be analysed using linear binary pattern. LBP approach used in [6] describes an efficient multimodal face recognition method. This method combines textured as well as depth features extracted from the input image. Linear polynomial approximation and[7] linear binary pattern methods are combined to extract the features and discrete Fourier transformation is used as a transformation tool. The effective face recognition method is achieved with the help of this technique.

III. Face Recognition using LDA

This approach [8] uses LDA and 2 channel wavelet transformation approach for face recognition. The 2 channel approach is used for factorization of half band polynomial. The analysed system also compares LDA approach with PCA. LDA approach for face recognition is described in this section.

[9]Linear Discriminant analysis is useful to determine combined features that do the separation of the classes. The length and complexity associated with the calculations are reduced using LDA approach. The dimensionality reduction and classification of face recognition is accomplished using least time and space complexity. Distortion within the image is common. This is also accomplished through LDA.

Mathematically, a set of n dimensional vectors x_i, x_2,-------,x_n belongs to l classes of faces.

Max $\frac{w^7 s_n w}{w^7 s_w w}$

Where

S_n=$\sum_{i-1}^{n} n_i(u^i - u_{total})(u^i - u_{total})$

S_w=$\sum_{i=1}^{n}\sum_{j=1}^{n}(x^i - u^i)(x^i - u^i)^i$

U is the mean of training images presented to the simulation. S_w is within the scatter matrix and $S_n$is between class scatter matrixes.

IV. Face Recognition using DWT

Face Recognition is critical in identification and verification of a person. This [10]approach using training images and applied with 2D-DWT to obtain LL band features. The LL band features are then normalized so that result lies between 01 and 1.The output obtained is compared against the original image to generate unique features. Gaussian filter [11] is applied in order to remove the noise if any within the image. Further, the feature vectors of many images are combined to form a unique feature vector representing several features. This process also performs compression and enhances recognition rate. The DWT approach is described in this section.

[12]DWT is widely used in numerical and functional analysis. In these areas wavelets are considered to be discretely distributed. DWT has advantage that both location and frequency information is considered. DWT has advantage over Fourier transformation since it has temporal resolution. The concept of wavelet is simple. They are used for multistage analysis process. Description of multistage wavelet is described considering the example as

Example 1

The sequence of wavelets are considered using n=2^3

y={1,1,2,3,1,3,2,2}

Consider vectors P and L computed through algorithm for multistage which can be applied as follows

1. $P_{I_{-1},_{J}} = \frac{1}{\sqrt{2}}(L_{I_{,2K}} - L_{I,2K_{-1}})$
2. $L_{I_{-1},_{J}} = \frac{1}{\sqrt{2}}\left(L_{I_{,2K}} + L_{I,2K_{-1}}\right)$
3. $I = I - 1$
4. If i=0 then stop else move to step 1

The basic ideas behind wavelets are portrayed through the above listed algorithm. The algorithm provides basic understanding of the wavelets or provides compact structure analysis of stored information.

The approaches described above have good recognition rate but it still can further improved. The proposed approach with SIFT key provides better results in terms of accuracy and recognition rate.

V. Proposed Technique for Face Recognition

The proposed scheme combines DWT and LDA method in which result is obtained through decomposition of metrics in four details sub bands. The information obtained is approximation details. The reduced image information is presented to PCA to obtain principal components and reduces dimensionality for storing. The proposed approach is capable of reducing image registration and is highly sensitive to skewing, pin cushioning and vignette that inevitably occurs in images. The proposed work takes the optimal properties of various algorithms along with genetic algorithm to produce optimal rate corresponding to face recognition. The proposed algorithm uses training dataset. The image is selected from training dataset. Median filtering is applied along with clipping operation in order to fetch only face part of the image. Feature extraction module is applied in order to fetch the features from the image. The invariant features are fetched from training set of faces. A feature is a selected image region with an associated descriptor. The descriptor is a special histogram of the image gradients. The gradient at each pixel is regarded as a elementary feature vector which is formed by the pixel location and their gradient orientation. After feature extraction, feature selection is applied. This is accomplished by the use of hybrid approach. In this approach population is initialized. The population consist of features which represent chromosomes. Each feature is fitted with the objective function. The selection process takes place through roulette wheel. In cross over phase chromosome gives rise to new generations. The mutation produces new chromosomes for better generations. In each round a set of similar chromosomes can be generated, these chromosomes may overlap the optimal features for extraction process. The best solution is taken into consideration. Identifying the non-dominated solution and sorting them in Preto front. The non-dominated solutions are copied to the next Pareto solution and the least crowded solutions are also added. The requirements when satisfied algorithm terminates.

The algorithm is listed as follows

Algorithm Face(Training_Set$_i$)

Input: A set of images describes as training set. Representation as training_set$_i$
Test_Image from Training_set$_{i.}$ Test_Image=Training_set$_i$

- Apply Face Acquisition and selection procedure to select particular test image from Training set.
- Apply median filter to reduce noise if any from the image

  Test_Image=median2(Test_Image)
- Clipping operation implementation to clip the image to extract only necessary portion of the image.
- Apply Feature extraction based upon descriptor,

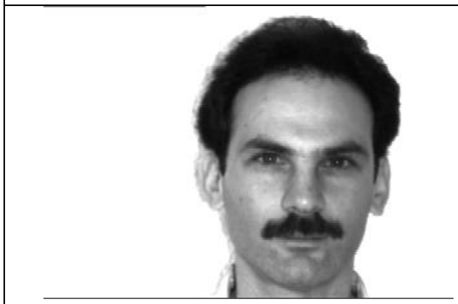
  Feature$_i$=Hist(Test_Image)
- Apply Hybridized optimization approach for optimum feature extraction

Repeat the above listed steps until termination criteria is satisfied or optimal result is obtained

VI. Results

The result produced with the proposed technique is compared with the other approaches. The recognition rate and accuracy is better with this approach as compared to existing approaches. The result in terms of optimal number of features selected is given through training images.

TABLE I. OPTIMAL NUMBER OF FEATURES SELECTED THROUGH PROPOSED APPROACH

Selected Image	Optimal Features selected
	62
	43
	95

The result generated in terms of accuracy is better as compared to existing approach. The accuracy is calculated in terms of error rate. The error rate is calculated considering the following formula

$$Error_{Rate} = ERF(Image_i)$$

$$Accuracy = 100 - Error_{Rate}$$

The Result generated is listed in the tabular structure.

TABLE II. ACCURACY OF VARIOUS CLASSIFIERS

Technique with Classifier	Accuracy(%)
PROPOSED	99.978
CC	99
DWT	99
DWT+CC	99
DWT+PCA	99

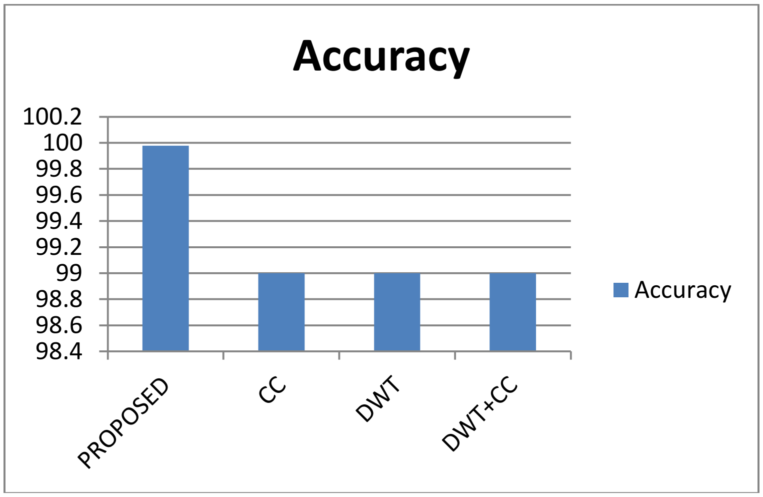


Fig. 1. Accuracy obtained with various classifiers

The optimal feature extracted from various classifiers is shown through tabular structure. The obtained features are best in case of proposed technique.

TABLE III. OBTAINED FEATURES FROM DIFFERENT CLASSIFIERS

Optimum feature selected	Proposed with SIFT	CC	DWT	DWT+CC	DWT+PCA
Training Set 1	95	85	93	92	90
Training Set 2	43	35	42	41	40
Training Set 3	93	82	90	89	88
Training Set 4	92	71	89	87	88
Training Set 5	90	80	88	85	86
Training Set 6	85	70	83	81	82
Training Set 7	82	72	78	75	77
Training Set 8	81	71	79	78	79
Training Set 9	87	72	84	82	85
Training Set 10	91	80	88	86	87

For demonstration 10 images from training set are used.

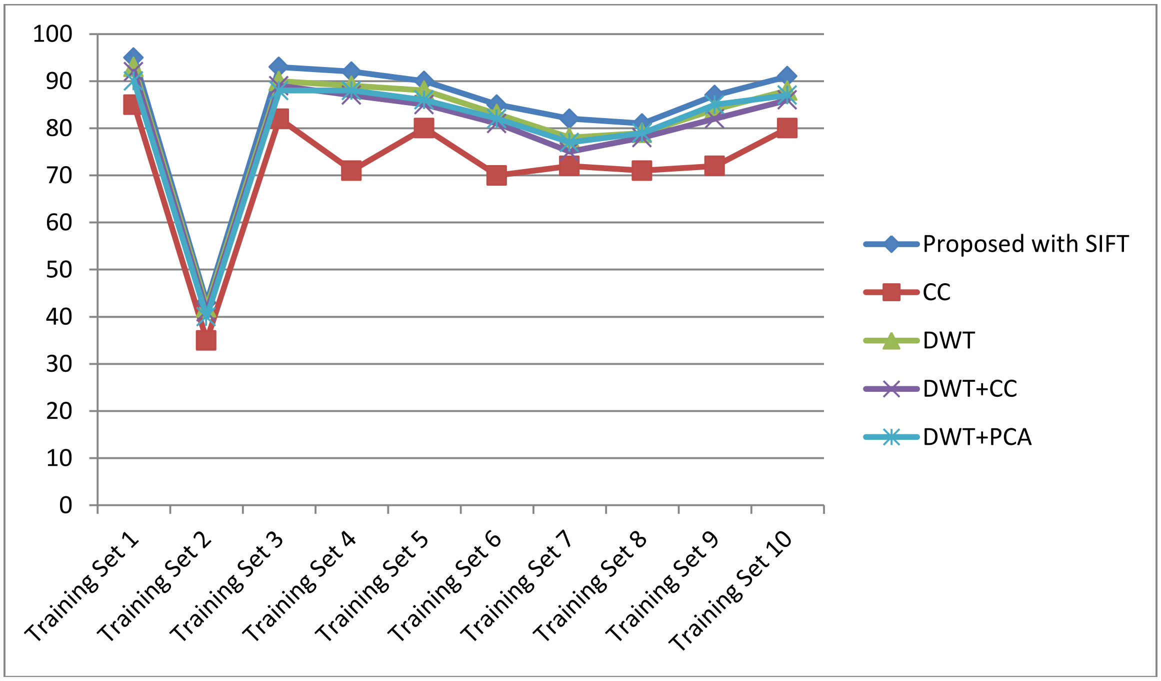


Fig. 2. Obtained features from different classifiers

The recognition rate obtained through proposed technique is better as compared to existing approach. This is obtained through simulation. Values are listed through tabular structure as

TABLE IV. RECOGNITION RATE THROUGH VARIOUS CLASSIFIERS

Training set	Proposed with SIFT	CC	DWT	DWT+PCA	DWT+CC
Training Set 1	98.6087	89.7801	93.0541	93.0401	89.8831
Training Set 2	96.0223	93.2043	94.6842	94.6662	93.1981
Training Set 3	72.0652	55.9279	64.2406	64.2447	57.6083
Training Set 4	92.0232	71.3455	89.2343	87.0873	88.6089
Training Set 5	90.1143	80.9850	88.890	85.8792	86.8769
Training Set 6	85.2304	70.5744	83.3424	81.0122	82.9034
Training Set 7	82.5304	72.4763	78.2344	75.3455	77.3423
Training Set 8	81.1203	71.2342	79.2342	78.3434	79.4555
Training Set 9	87.5607	72.3443	84.2342	82.8237	85.2332
Training Set 10	91.5871	80.3434	88.3453	86.7878	87.0452

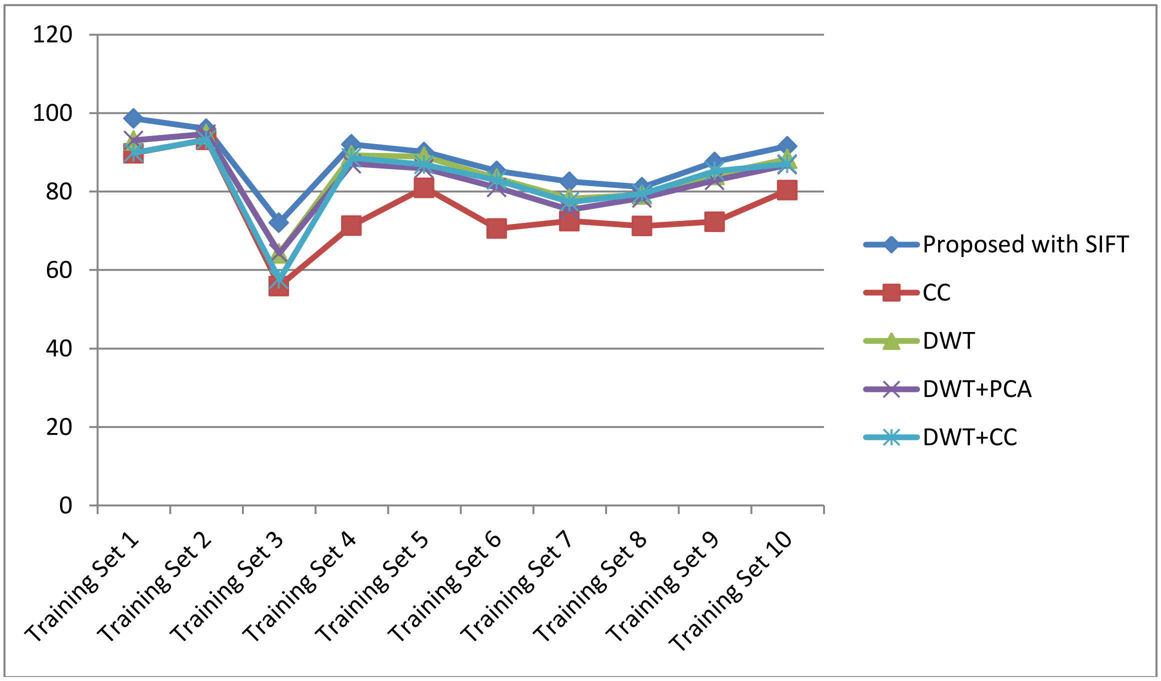


Fig. 3. Recognition rate through various classifiers

Results shows that proposed technique is better in terms of number of different characteristics. These characteristics include recognition rate, accuracy, optimal number of features and obtained features through classifiers.

VII. CONCLUSION AND FUTURE SCOPE

Automatic selection of features is greatest advantages of the proposed technique. The feature selected is fed into the system for optimality. Since the process is iterated hence result obtained is filtered with iteration. The results are compared against different classifiers to prove the validity of the approach. The feature selection is based on objective function value. Better convergence in terms of recognition rate is presented through proposed technique. The result is better in terms of recognition rate, accuracy and number of optimal features generated.

The proposed technique utilizes hybrid approach to obtain better convergence in terms of recognition rate. The genetic algorithm which is used is situation dependent. By saying so we mean it may converge better for some of training sets better as compared to other training sets. Unguided mutation is also problem with the proposed genetic approach. In future hybrid approach of Ant colony and Honey bee can be used for achieving enhanced optimality.

REFERENCES

[1] "What is facial recognition? - Definition from WhatIs.com." [Online]. Available: http://whatis.techtarget.com/definition/facial-recognition. [Accessed: 29-Dec-2016].

[2] T. kaur and R. dhir, C. Science," Hexagonal Descriptor Particle Swarm Optimization with Knowledge-Crowding for Face Recognition",SERSC, EI compendex, IJSIP vol. 9, pp. 253–264, 2016.

[3] T. Pecchia, A. Gagliardo, C. Filaninno, P. Ioalè, and G. Vallortigara, "Metadata of the chapter that will be visualized in SpringerLink Adult-Born Neurons in the Olfactory," Behav. Lateralization Vertebr., 2012.

[4] P. Adibi and S. Ahmadkhani, "Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework," IET Comput. Vis., vol. 10, no. 3, pp. 193–201, Apr. 2016.

[5] H.-W. Guo, Y.-S. Huang, C.-H. Lin, J.-C. Chien, K. Haraikawa, and J.-S. Shieh, "Heart Rate Variability Signal Features for Emotion Recognition by Using Principal Component Analysis and Support Vectors Machine," in 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE), 2016, pp. 274–277.

[6] M. C. Sobia, V. Brindha, and A. Abudhahir, “Facial expression recognition using PCA based interface for wheelchair,” in 2014 International Conference on Electronics and Communication Systems (ICECS), 2014, pp. 1–6.

[7] Naveen S., S. S. Nair, and R. S. Moni, “3D face recognition using optimised directional faces and fourier transform,” in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015, pp. 1856–1861.

[8] J. Ren, X. Jiang, and J. Yuan, “LBP-Structure Optimization With Symmetry and Uniformity Regularizations for Scene Classification,” IEEE Signal Process. Lett., vol. 24, no. 1, pp. 37–41, Jan. 2017.

[9] M. A. Muqeet and R. S. Holambe, “Face recognition using LDA based generalized half band polynomial wavelet filter bank,” in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016, pp. 4649–4653.

[10] D. Marculescu, N. H. Zamora, P. Stanley-marbell, and R. Marculescu, “Fault-Tolerant Techniques for Ambient Intelligent Distributed Systems-,” pp. 348–355, 2003.

[11] G. V Sagar, S. Y. Barker, K. B. Raja, K. S. Babu, and Venugopal K R, “Convolution based Face Recognition using DWT and feature vector compression,” in 2015 Third International Conference on Image Information Processing (ICIIP), 2015, pp. 444–449.

[12] Singh, “Performance Comparison of Various Image Denoising Filters Under Spatial Domain,” vol. 96, no. 19, pp. 21–30, 2014.

[13] Xue B. Zhang M., Browne W. N., “Particle Swarm Optimization for Feature Selection in classification: A Multi-Objective Approach”, IEEE Transactions on Cybernetics, pp. 1-16, 2012

[14] T. Kaur and R. Dhir, "Feature Optimility Based Semi-Supervised Face Recognition Technique" ,chapter in Advs in Intelligent Syst., Computing, Vol. 515, proceeding of the 5th international conference on frontiers in intelligent computing.

[15] T. Kaur and R. Dhir, “A Non-domination Pareto-based Scale-Invariant Approach for Face Recognition,” Eur. J. Eng. Res. Sci., vol. 1, no. 2, pp. 6–13, 2016.

Empirical Evaluation of Social and Traditional Search Tools for Adhoc Information Retrieval

Safdar Hussain
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Nadeem Akhtar
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Intesab Hussain
Dept. of Computer System Eng.
Quaid e Awam Univ.
Nawabshah

Malik Muhammad Saad Missen
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Mujtaba Husnain
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

M. Ali Nizamani
Faculty of Engineering, Science and Tech., ISRA Univ.
Haiderabad

***Abstract*—The nature of World Wide Web (www) has evolved over the passage of time. Easier and faster availability of Internet has given rise to huge volumes of data available online. Another cause of huge volumes of data is the emergence of online social networks (like Facebook, Twitter, etc.) which has actually changed the role of data consumers to data generators. Increasing popularity of these online social networks has also changed the way different web services used to be used. For example, Facebook messaging has some impact on usage of emails; twitter usage affects (positively or negatively) online newspaper readings. Both of these platforms are heavily used for information searching. In this paper, we evaluate the role of Facebook and Twitter for academic queries and compare the findings with Google search engines to find out if there is a chance that these online social networks will replace Google sooner. A query set selected from the standard AOL dataset is used for experimentation. Academic related queries are selected and classified by expert users. Findings of Google, Facebook and Twitter are compared against these queries using Mean Average Precision (MAP), as a metrics for evaluation. Results conclude that Google has the dominating factor with a better MAP than Facebook and Twitter.**

Keywords—AOL Query Log; Facebook; Twitter; Social Search

I. INTRODUCTION

With the passage of time, the nature of web has evolved. A major breakthrough in this regards is the emergence of online social networks. Online social networks have not only changed the role of users from content consumers to content generators but also have changed the way users used to search the web. These social media websites, showing various forms of consumer generated content (CGC) such as virtual communities, blogs, social networks, wikis, collaborative tagging and media files that are shared on sites like Flickr and YouTube have gained substantial popularity [1]. Also Social network sites (SNSs) such as Facebook, MySpace, Bebo, and Cyworld have attracted millions of users, many of whom have assimilated these sites into their daily practices in real time [2]. Apart of it, most sites support the maintenance of pre-existing social networks, but others help strangers connect based on shared interests, activities, or political views. Number of sites gratify to diverse audiences, while others attract people based on shared racial or common language, religious, sexual, or nationality-based identities. These sites also vary in extent to which they incorporate new communication and information tools, such as blogging, mobile connectivity, and photo/video-sharing [3]. This exponentially increasing interest in online social networks (see figure 1) has resulted in generation of huge amount of daily data on the web. Traditionally, it is know that search engines are used for searching relevant information from the web. However, there has been an increasing trend of searching information using online social networks. This is where the concept of social search gets emerged.

A. Social Search

The process of social search on social media points out the usage of social mechanism to seek information on web. Many search engines provide facility for social search; by providing a link of a web page (e.g., public Twitter posts), or it is simply a process of result ranking [4]. Social tagging systems' output can be the base platform for online social search engines like delicious on (delicious.com). Evan et al. [6] point out the stages for search process in cases when people' need to be in contact with others. Morris et al. [7] provide a survey for Twitter and Facebook users for the cases; to have a status message question type about any social networks need. The study of Social searching behavior, on a Q&A site, is to post a question (e.g., Harper et al. [8], Liu et al. [9]) on community of large scale users (normally having no direct relation to the asker) can put answers. The systems like Aardvark which is simply a system of expertise-finding [10] or Collabio [[11]], straightaway can be useful to help in person finding process, and is qualified for information need consideration. Reference librarians can provide assistance as professionals to numerous searchers [12]. The social search asserts that (a) social network links can be leveraged to improve the quality of search results, and that (b) a growing body of Internet content cannot be retrieved by traditional web search as it is not well-connected to the hyperlinked Web[14], [13]. It is said that current web search engines are not able to find relevant information available on online social networks. Therefore, there is a trend of using online social networks for information seeking. In this work, we focus to analyze this trend by looking at the relevancy of the results both kind of search tools return. We try to find out how much successful are online social networks on providing relevant results and if there is any chance of online social networks replacing traditional search engines. We select a set of academic queries for this purpose because academic queries are one of the most searched information on online social networks. The overall prime objective of this work is to compare and evaluate the effectiveness of online social networks and traditional search engines for search of academic queries. A diversity of topics is selected from a standard query log that relates to different academic information needs.

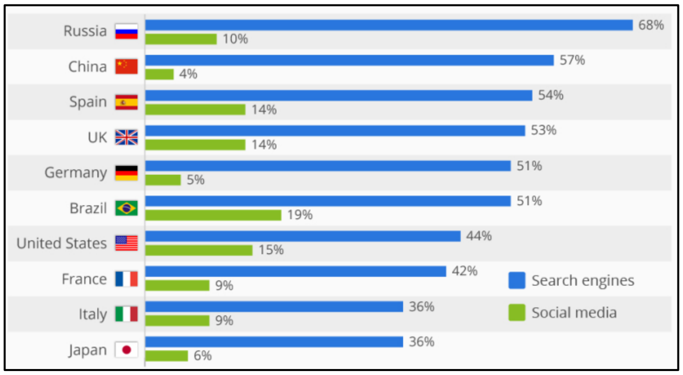


Fig. 1. Search Engines Vs Online Social Networks

The paper is categorized in different sections like: Section II, contains some related literature work while section III portrays experiments and discuss their results. At the end, we conclude our paper with conclusions drawn from our work.

II. Related Work

Most of the other works typically focus on social search. For example, Dodds et al. [16] report a successful experiment on exploration of social search. Experiments performed in which more than 60,000 email users attempted to reach one of the 18 target persons in 13 countries by forwarding messages to acquaintances. It was found that targets can be grasped in a median of 5 to 7 steps. Another work which tries to improve web search using social aspects is done by Bao et al. [17]. They used social annotations for this purpose. Some have also analyzed the impact of users social networks on personalization [18]. There are works that have evaluated some specific online social network and evaluated them for social search. For example, Scale et al. [21] evaluate the role of Facebook platform as a social search engine. They found out that Facebook returns irrelevant results for unknown persons or groups. Another very popular work in this regard has been performed by Tancer [22]. Tancer put in front a case study of a user information need, the solution in which is delivered by friends in Facebook relieving the users' use of a traditional search engine. Tancers experience concludes insight in how humans in a Social Networking Sites (SNS) environment can collaborate and participate to meet user information needs. One of the most important contributions towards social search is proposal of models for social search. Work of Evan et al. [19], [20] is considered a significant effort in this regard. According to Jaime Teevan et al. [23], roundabout 50% users are in contact via the use of Status Message Question Asking(SMQA) behavior, so that is the reason that SMQA is the hot research area and most common item in new researches. After 50% Facebook users, twitter was on second with 33% and LinkedIn, Google with 25% on third in usage of SMQA. There are some works that we find very relevant to our work in nature of the problem they worked on. The work of Morris et al. is one of the initial works [15] focusing on social search. This work is most related to our work however there are major differences between our methodology and target domains. Compared to our work where we effectively use SNs online search option, they used status messages as information seeking option. Similarly, Zheng et al. [24] tried to evaluate online social networks for travel queries. The focus and goal of their study was to examine the extent to which social media results appear in search engine results in the context of travel-related searches. Their employed research design simulated a traveler's use of a search engine; it was for travel planning by using a set of pre-defined keywords in combination with nine U.S. tourist destination names. Comparative findings of search results reveal that the role play of social media contains significant portion of the search results, as now people' rely & use social media community more than ever before. The current work is the confirmation to argue that social media provides online search progressively. Another work that we find somehow close to our work is done by Alan et al. [25]. In current paper, authors examined the work potential for using online social networks to boost Internet search. They analyzed the differences between the social networking systems and Web in terms of the mechanisms they use to locate and publish useful information. They conferred the benefits of integrating the mechanisms for finding useful content in both the social networks and Web. Such initial results from a social networking experiment suggest that such integration has the potential to improve the quality of Web search experience. Our work portrays the results by evaluating the situations in which platforms are suitable for what type of categories on different platforms like Google, Facebook & Twitter.

III. Experiments

Our experimentation significantly follows standard methods and measures. Experiments start with selection of academic queries [26]. We consider a query an academic query if it seeks any information relating to academics needs (for example, from admission information search to expert searching). We use real world search engine query log for extraction of academic queries.

A. *Selection of Queries*

On August 4, 2006 (in first decade of century), AOL (America Online) intuitively released a huge dataset query log collection (i.e. of 500,000 people') that was the collection of real users search relation with AOL for academic (non-commercial) domain. AOL, take an action and immediately (on August 7, 2006) cleared the site with such data, but it was too late. The files were floated and shared all over the internet within this short time span. It was bulk of about 36 million Web searched queries typed by approximately 657,000 users for three month time span (from March 01, 2006, to May 31, 2006). It consisted of a compressed 439 MB download with 2.12 GB in expansion. A sequential go through AOL dataset makes it possible for us to select a subset of academic queries. We identify and further categorized these queries under different information need labels to make them more understandable. Table1 gives 36 queries for 6different information needs are given in the table below.

TABLE. I. CATEGORIES AND THEIR QUERIES

Sr. No	Categories	Queries
1	Research Papers	• how to write research paper introduction thesis hypothesis, • thesis statements research papers, • technical writing research paper tips, • how to cite your information in your research paper, • bibliography for research paper, • structure of research paper
2	Distance Learning	• University that offers PhD program from distance learning, • Pros & cons for distance learning in high school, • Distance learning education council, • Army distance learning, • Distance learning undergraduate degree universities, • Virtual classes distance learning online courses.
3	Research Topics	• What is meant by Educational Research Topics? Enlist them and explain, • What is the process for choosing a research topic? Discuss some key points? • Name the topics for a good research paper. Briefly tell about all of them as well, • What are current research topics? Specifically in computer science, • How to perform qualitative research on any topic? Explain, • Terrorism topic for research paper.
4	Scholarship Program	• How to get easy scholarships for computer science PhD program? • What are 21st century scholarship programs? • What do you know about Bill Gates scholarship program? Comment, • What is National Merit Scholarship Program? • What is Microsoft scholarship program? • What is California state scholarship program?
5	University Programs	• What are Colorado technical university computer science programs? • What is New York university summer program? Any detail, • What are Columbia university PhD programs? Any detail, • What are university travel study programs? Comment about it, • What is Oxford university summer program? Any detail, • What are Texas southern university PhD programs? Comment about it.
6	Research Institutes	• What is the Christian research institute? • What are National research institutions in Pakistan? Write detail, • Electric power research institute in Pakistan, • Southwest research institute, • What is the role of Economic cycle research institute? Comment, • What is Virtual research institute? Describe about it.

The queries given in Table 1 are used to compare search engines with online social networks. We select three different platforms for performing our experiments. We choose Google to represent Search Engines while Facebook and Twitter are chosen for representing online social networks. This selection is based on the popularity of each platform (see figure 1) that can be used for textual information search.

B. *Returned Results and Evaluations*

In next phase of experimentation, we use search interface of each selected platform to search with selected list of queries (see table 1). Top 20 documents for each query are downloaded for each selected platform as it is shown in figures 2, 3 and 4. To evaluate these returned results, a massive exercise of user evaluation is planned.

C. *User Evaluations*

We recruit five different users for unbiased evaluation of returned results foreach platform. Each user is aged between 24 to 30 years and is computer science graduates. Users are asked to thoroughly understand the queries for unbiased evaluation of returned results. They are presented with a web interfaceto mark each returned result as relevant (1) or not relevant (0). Evaluationsare performed in a sequential process i.e. first of all results of all queries areevaluated for Google and then same process is repeated for Twitter and Facebook. Fleiss kappa [27] is used to measure inter-annotator agreement for eachselected platform as shown in table 2. We can see that results are good enough to be considered as a reliable inter-annotator agreement.

TABLE. II. MAP FOR GOOGLE PLATFORM

Queries	**Users**				
	User-1	User-2	User-3	User-4	User-5
Category – I (Research Papers)					
Q1	0.82	0.88	0.94	0.89	0.90
Q2	0.88	0.84	0.94	0.94	0.94
Q3	0.73	0.80	0.76	0.79	0.80
Q4	1	1	1	1	1
Q5	1	0.95	1	1	1
Q6	1	1	1	1	1
MAP/User	0.91	0.91	0.94	0.94	0.94
Category – II (Distance Learning)					
Q7	0.95	0.94	0.95	0.95	0.95
Q8	0.72	0.63	0.78	0.79	0.84
Q9	1	1	1	1	1
Q10	0.74	0.83	0.82	0.82	0.84
Q11	0.95	1	0.95	0.95	0.95
Q12	0.85	0.94	0.89	0.90	0.90
MAP/User	0.87	0.89	0.90	0.90	0.91
Category – III (Research Topics)					
Q13	0.73	0.67	0.79	0.80	0.80
Q14	0.90	0.90	0.90	0.90	0.90
Q15	0.79	0.75	0.88	0.84	0.85
Q16	0.89	0.89	0.90	0.90	0.90
Q17	0.83	0.88	0.85	0.85	0.85
Q18	0.88	0.89	0.88	0.89	0.90
MAP/User	**0.84**	**0.83**	**0.87**	**0.86**	**0.87**
Category – IV (Scholarship Program)					
Q19	1	1	1	1	1
Q20	1	1	1	1	1
Q21	0.89	0.90	0.89	0.89	0.90
Q22	1	1	1	1	1
Q23	0.88	0.88	0.89	0.89	0.89
Q24	0.94	0.94	0.95	0.95	0.95
MAP/User	**0.95**	**0.95**	**0.96**	**0.96**	**0.96**
Category – V (University Programs)					
Q25	0.94	0.95	0.89	0.89	0.89
Q26	0.90	0.95	0.90	0.90	0.90
Q27	1	1	1	1	1
Q28	1	1	1	1	1
Q29	0.95	0.90	0.95	0.95	0.95
Q30	1	1	1	1	1
MAP/User	**0.97**	**0.97**	**0.96**	**0.96**	**0.96**
Category – VI (Research Institutes)					
Q31	0.94	0.95	0.89	0.90	0.90
Q32	0.95	0.95	0.90	0.90	0.90
Q33	0.89	0.89	0.94	0.95	0.95
Q34	1	1	1	1	1
Q35	1	1	1	1	1
Q36	0.83	0.83	0.85	0.85	0.85
MAP/User	**0.94**	**0.94**	**0.93**	**0.93**	**0.93**

TABLE. III. COMPARISON OF MAPS FOR ALL PLATFORMS (GOOGLE, TWITTER, FACEBOOK)

		Google	**Twitter**	**Facebook**
Category – I	**Q1**	0.89	0.49	0.65
	Q2	0.91	0.45	0.65
	Q3	0.78	0.79	0.47
	Q4	1	0.95	0.42
	Q5	0.99	0.39	0.31
	Q6	1	0.35	0.12
MAP Per Category		**0.93**	**0.57**	**0.44**
Category –I I	**Q7**	0.95	1	0.63
	Q8	0.75	0.70	0.74
	Q9	1	0.69	0.74
	Q10	0.81	0.05	0.70
	Q11	0.96	0.67	0.56
	Q12	0.90	0.63	0.76
MAP Per Category		**0.90**	**0.62**	**0.69**
Category –III	**Q13**	0.76	0.42	0.40
	Q14	0.90	0.55	0.61
	Q15	0.82	0.48	0.66
	Q16	0.90	0.65	0.55
	Q17	0.85	0.25	0.50
	Q18	0.89	0.33	0.58
MAP Per Category		**0.85**	**0.45**	**0.55**
Category –IV	**Q19**	1	0.72	0.36
	Q20	1	0	0.53
	Q21	0.90	0.47	0.55
	Q22	1	0	0.54
	Q23	0.89	0.71	0.42
	Q24	0.95	1	0.48
MAP Per Category		**0.96**	**0.48**	**0.48**
Category –V	**Q25**	0.92	0	0.66
	Q26	0.91	0.58	0.51
	Q27	1	0.61	0.56
	Q28	1	0.68	0.50
	Q29	0.94	0.56	0.54
	Q30	1	0	0.43
MAP Per Category		**0.96**	**0.40**	**0.53**
Category –VI	**Q31**	0.92	0.44	0.62
	Q32	0.92	0.53	0.60
	Q33	0.92	0.40	0.65
	Q34	1	0.37	0.49
	Q35	1	0.54	0.50
	Q36	0.84	0.38	0.61
MAP Per Category		**0.93**	**0.44**	**0.58**

We decided the usage of mean average precision (MAP) [28] as metric for performance evaluation of each platform. In a set of queries, the MAP is the mean of the average precision scores for each query.

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

Where Q is the total number of queries and AP(q) is average precision for a given query q.

To compute average precision, it is assumed that we have total twenty relevant documents in the collection for each query. Following tables provide MAP for each selected platform computed through labeling by each user. Looking at individual results for Google (figure 2), Facebook and Twitter, it can be concluded that MAP results for Google are the highest and consistent across different categories as well as different users. MAP values for Twitter are much lower but consistent across different categories and users. However, for Facebook results we see inconsistency among users as well as among categories. Comparing MAP results for all three platforms using figure 2, we can conclude that Google has produced the best results for all academic queries while Facebook has beaten Twitter for most of the categories.

TABLE. IV. FLEISS KAPPA FOR DIFFERENT PLATFORMS

Sr. No	Platform	Fleiss Kappa
1	Google	0.79
2	Twitter	0.70
3	Facebook	0.73

TABLE. V. VARIANCE COMPARISONS

Sr. No	Platform	Variance Among Users	Variance Among Categories
1	Google	0.0001	0.001
2	Twitter	0.001	0.007
3	Facebook	0.004	0.007

Table 5 shows the variance among MAP for users and also for categories which also show that Google result show more consistent attitude for all query types. Therefore, it can be concluded from all results that Google still holds its position for academic information searching. However, there is a trend of seeking support of online social networks for search information which did not exist earlier. We also observed that Facebook proved to be more helpful when searching for academic related information than Twitter. Main reason for these results is presence of many Facebook pages and groups that share much academic related information such as admissions and scholarship opportunities.

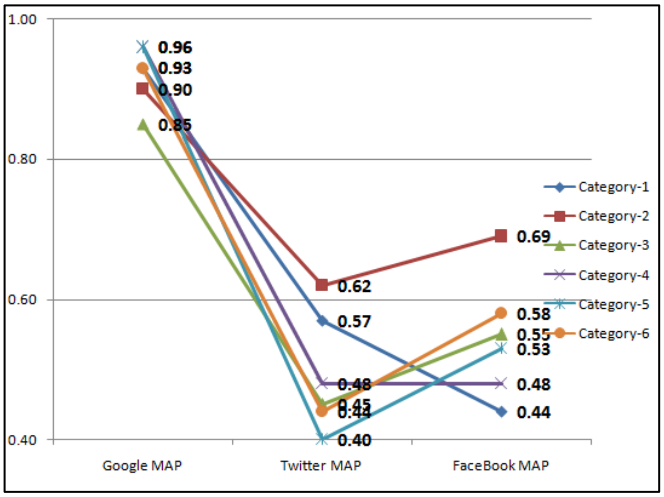


Fig. 2. MAP Comparison of Google, Twitter and Facebook for all Query categories

IV. CONCLUSION

In this paper, we made an effort to compare social search with traditional search for academic queries. The main objective was to evaluate who is better after years of dominance of online social networks among web users. For this purpose, we selected Facebook and Twitter for representing online social networks while Google search engine is used for representation of traditional search. We used AOL data-set for selection of queries. The experimentation results reveal that Google maintains its dominance in academic information searching. Comparing both Facebook and Twitter, it has been found that Facebook provides much more relevant information for academic queries to its users than Twitter.

REFERENCES

[1] Lewandowski, Dirk, ed. Web search engine research. Emerald Group Publishing Limited, 2012.

[2] Ramage, Magnus. Online communication and collaboration: A reader. Routledge, 2010.

[3] Ellison, Nicole B. "Social network sites: Definition, history, and scholarship." Journal of Computer‐Mediated Communication 13.1 (2007): 210-230.

[4] Smyth, Barry, et al. "Google shared. a case-study in social search." International Conference on User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, 2009.

[5] Morris, Meredith Ringel, Jaime Teevan, and Steve Bush. "Enhancing collaborative web search with personalization: groupization, smart splitting, and group hit-highlighting." Proceedings of the 2008 ACM conference on Computer supported cooperative work. ACM, 2008.

[6] Evans, Brynn M., and Ed H. Chi. "Towards a model of understanding social search." Proceedings of the 2008 ACM conference on Computer supported cooperative work. ACM, 2008.

[7] Morris, Meredith Ringel, Jaime Teevan, and Katrina Panovich. "What do people ask their social networks, and why?: a survey study of status

message q&a behavior." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2010.

[8] Harper, F. Maxwell, et al. "Predictors of answer quality in online Q&A sites." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2008..

[9] Liu, Yandong, Jiang Bian, and Eugene Agichtein. "Predicting information seeker satisfaction in community question answering." Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008.

[10] Horowitz, Damon, and Sepandar D. Kamvar. "The anatomy of a large-scale social search engine." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[11] Bernstein, Michael, et al. "Collabio: a game for annotating people within social networks." Proceedings of the 22nd annual ACM symposium on User interface software and technology. ACM, 2009.

[12] Taylor, Robert S. "Question-negotiation and information seeking in libraries." College & research libraries 29.3 (1968): 178-194.

[13] Nextmedia, C. S. A. "Social networks overview: Current trends and research challenges." European Commission Information Society and Media (2010).

[14] Tsou, Ming-Hsiang, et al. "Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election." Cartography and Geographic Information Science 40.4 (2013): 337-348.

[15] Morris, Meredith Ringel, Jaime Teevan, and Katrina Panovich. "A Comparison of Information Seeking Using Search Engines and Social Networks." ICWSM 10 (2010): 23-26..

[16] Dodds, Peter Sheridan, Roby Muhamad, and Duncan J. Watts. "An experimental study of search in global social networks." science 301.5634 (2003): 827-829.

[17] Bao, Shenghua, et al. "Optimizing web search using social annotations." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[18] Carmel, David, et al. "Personalized social search based on the user's social network." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[19] Evans, Brynn M., and Ed H. Chi. "An elaborated model of social search." Information Processing & Management 46.6 (2010): 656-678.

[20] Evans, Brynn M., and Ed H. Chi. "Towards a model of understanding social search." Proceedings of the 2008 ACM conference on Computer supported cooperative work. ACM, 2008.

[21] Scale, Mark-Shane. "Facebook as a social search engine and the implications for libraries in the twenty-first century." Library Hi Tech 26.4 (2008): 540-556.

[22] Tancer, Bill. "Is Facebook the future of search?." Time magazine (2008).

[23] Oeldorf-Hirsch, Anne, et al. "To search or to ask: the routing of information needs between traditional search engines and social networks." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014.

[24] Xiang, Zheng, and Ulrike Gretzel. "Role of social media in online travel information search." Tourism management 31.2 (2010): 179-188.

[25] Mislove, Alan, Krishna P. Gummadi, and Peter Druschel. "Exploiting social networks for internet search." 5th Workshop on Hot Topics in Networks (HotNets06). Citeseer. 2006.

[26] Ellis, David. "Modeling the information-seeking patterns of academic researchers: A grounded theory approach." The Library Quarterly 63.4 (1993): 469-486.

[27] Fleiss, Joseph L. "Measuring nominal scale agreement among many raters." Psychological bulletin 76.5 (1971): 378.

[28] Zhu, Mu. "Recall, precision and average precision." Department of Statistics and Actuarial Science, University of Waterloo, Waterloo 2 (2004): 30.

Mobile Technology based Polio-Vaccination System (PVS) – First Step Towards Polio-Free Pakistan

Nukhba Afzal
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Amnah Firdous
Dept. of Computer Science
CIIT,
Vehari

Hina Asmat
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Malik Muhammad Saad Missen
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Nadeem Akhtar
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur,
Pakistan

Saleem Ullah
Dept. of Information Tech
KFUEIT,
Rahimyar Khan

***Abstract*—Health information technology revolutionized the world with its great expansion and widespread in the domain of health care system. Most of the developed countries adopted advanced technology in their vaccination systems. Vaccination systems of many developing countries still lack the use of technology eventually causing mismanagement and corruption to occur in vaccination campaigns. Issues like mismanagement and corruption not only affect vaccination campaigns but also cause further diffusion of a disease. Pakistan is also one of such countries where vaccination system is prone to these and many other issues and hence it does not help in disease eradication. For example, polio remains alive in Pakistan because Pakistan's Polio vaccination system is faced with many problems and the biggest one is security of vaccination teams. Corruption, mismanagement, unawareness among public and life-threat to vaccination teams are the main problems of current polio vaccination system of Pakistan. To overcome these flaws and to make an idyllic system with the new advanced technology, we propose technology oriented secure polio vaccination system. The proposed system is more secure and removes flaws in the current system. We model our proposed system using Colored Petri Nets (CPNs) which is a state-of-the-art tool for formal modeling.**

Keywords—Polio Vaccination; Information system; GPS technology; Health-care

I. Introduction

Advancement in technologies has led to the development of health care technologies. Technology in health care encourages the people to adopt healthy life style along with the advancement of technology-based interventions made to improve the working of health-care like remotely access the doctors, , clinical management support, supporting clinical diagnosis and treatment, send patients diagnosed results on time through SMS based appointment reminders, timely response. Thus there is a need to adapt technology for health-care system especially for under-developed countries and especially for the infectious diseases. Polio is one of the infection disease caused by poliovirus and still spreads only in Pakistan, Afghanistan and Nigeria [1, 2, 3]. Existing polio vaccination system of Pakistan is incompetent and ineffective. Mismanagement, corruption and insecurity are major problems of the existing system. These lacking take the vaccination campaign towards the termination and increase the rate of missed children. Therefore, there is a need of such a system which can cope with current challenges and have a capability to cope up with the current time and technology.

Polio remains endemic in two countries of the world – Afghanistan as well as Pakistan. Polio endemic countries could be the reason of importing wild polio virus to those countries that are non-polio endemic [5].Some measures have been taken to stop this spread of polio virus to other nations like some countries who are not endemic already declared polio vaccination mandatory for endemic countries travelers, who are not immunized against polio. Besides this, health care providers have identified the technologies with different purposes like in educating the people, diagnosing the diseases; management and communication between patient and health service providers to further cope with this infectious disease. However, immunization is considered to be one of the greatest health interventions to prevent polio. There is a need of making the process of immunization more effective by leveraging the support of advance technology. Existing health care systems with integration of new mobile technology can help to eradicate polio from world map [4, 6, 7].

In Pakistan, polio immunization campaigns have been facing nonstop setbacks. According to 2015 report of World Health Organization (WHO), 51 confirmed cases of polio virus occurred in Pakistan. There are many reasons behind having Polio in Pakistan that include lack of education and awareness, religious concerns, mismanagement, corruption and life threats to vaccinators.

In this paper, we propose a very effective technology based polio vaccination system for Pakistan after identifying the drawbacks in the current system. The main objective is to make the polio vaccination system efficient in general and security perspective. Without proper security of health workers and general public it is not possible to carry out a successful campaign [3][1].

A. Research Objectives

Our aim is to rectify the current polio vaccination system of Pakistan by making an effective use of technologies:

- To perform the detail study of current polio vaccination system of Pakistan to identify the problems and flaws (general as well as security related) in the current polio vaccination system,
- To propose a new technology based and secure polio vaccination system,
- To formally model our proposed system using formal modeling techniques.

II. RELATED WORK

We discuss related work with respect to countries that have their polio vaccination working.

A. United States

Technology is widespread in US with its great expansion. This extensive and prevalent usage of advanced tools gives an idyllic and perfect platform in order to help in the deliverance of vaccines.

The presence of electronic health record utilizing and the resultant capability to have vaccination data in E-record form grant an imperative establishment for delivering IT based vaccine interventions. Vaccination system become rationalized with the advanced tools like use of E-data, increase the number of patients interacted and contacted with the intervention [3, 8].

In United States, National vaccine recommendations goal is to increase the amount of vaccine-preventable diseases for diminution, abolition, or abolition. One of their objectives is to check out the rate of vaccination before the implementation of suggested vaccination and its preventable diseases with the comparison of transience and death rate [8, 9].

Electronic data made the work easier in keeping the accurate and reliable data. Whenever data is needed, there is no need to check out or find out the manual data like people do before the advancement of the advanced tools and technology. Computerized registries are being done now days, which help to sustain the record with its confidential. Once electronic record has been created, it can be used repeatedly or can be used by any other organizations as well like public organizations or any semi government organizations. Now a day in health care system, management can easily maintain the record of infants or adults. Members of health provider could easily fetch out the data of patient anytime or anywhere and its relevant record [3].

B. Nigeria

Nigeria has made extraordinary growth against wild poliovirus. World Health Organization (WHO) publicized that polio virus is not any more in Nigeria in September 2015.No case of wild poliovirus has reported in Nigeria from 2014. For the very first instance, Nigeria has broken up importation of wild polio virus, taking the nation state and the entire African territory nearer than ever to being experienced as free polio virus in the world.

Recently in 2012, most of the polio virus cases reported in Nigeria globally. Ever since, an intensive attempt by the management of Nigeria, religious leaders, civil society and thousands of enthusiastic vaccinators of polio virus have resulted in Nigeria productively bring it to halt. Not only had this but also improving vaccination coverage and adoption of new advanced technology in vaccination coverage also helped in reaching the goal.

Nigeria changed their way of primitive coverage of vaccination by adopted advanced tactics. They identified the programmatic weakness in polio campaign, tried to improve training, planning, supervision and accountability. They worked on global positioning system (GPS) device to improve the coverage and to easily track the vaccination teams.

Currently, tracking of different teams of vaccinators with GPS is carried of each campaign wise. A device is given to each team of vaccinator before the initiation of their work. As they finished their per day work assignment they again give a call to the head quarter about the finishing of their daily bases job. After taking the devices back from the each team of vaccinators, they call up for the daily progress report and reviews. So, this workflow is followed till the accomplished of vaccination process [9].

Such a project with Graphical information system technology improved the quality of micro planning and provided information about team performance. It also provided tools in identified missed children. Different branch to the Geographical Information System toil in Nigeria:

1) To targeting the different states of Nigeria with the GPS based digital map, it gathers the data with its relative boundary.

2) Incorporating of digital the map into the micro planning procedure,

3) A group of people has been bounded to keep eyes on the vaccinator along with their teams, with GPS tracker to chase them and to reach them with the help pg GPS navigation or the digital map.

4) Giving timely based reviews and feedback.

5) A web based portal has been set in with the GPS tracking system to view working of vaccination teams. So

[1] Poliomyelitis Fact sheet N°114". WHO int. October 2014. Retrieved 3 November 2014

management can check out the way campaign conducted at local level or remote level [9].

The use of digital map instead of old hand drawn map improves the campaign. Before that, hand sketched map were used, which were often incomplete and inaccurate. Small region or hamlet areas missed in hand drawn map and only major cities were located. Hamlet areas are more crucial in providing immunization. Identification of the village areas is tricky, while they are an artificial erect based on immediacy. Digital maps are printed for each and every zone for use in micro planning. An outline is used for the micro plan, together with the maps, is also made accessible via web based portals. The maps and the given outline is used by the each zone team. According to the area or the population, these essentials are given to each vaccination teams till the completion of vaccination campaign.

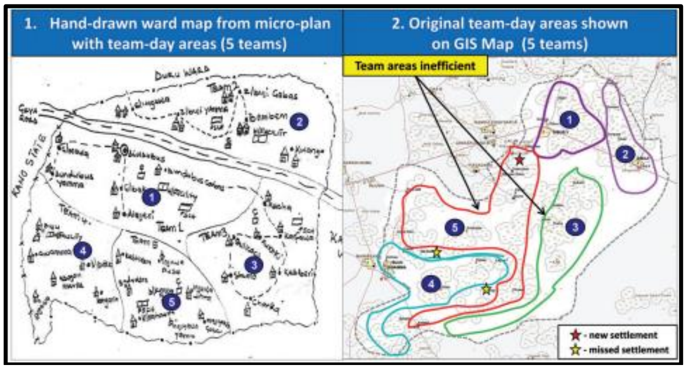

1. Hand-drawn ward map from micro-plan with team-day areas (5 teams)
2. Original team-day areas shown on GIS Map (5 teams)
Team areas inefficient
1
2
3
4
5
★ - new settlement
☆ - missed settlement

Fig. 1. Human hand drawn and geographical system maps

Nigeria did their best to completely eradicate the polio virus from their country. They used the new advanced technology and other tools which help them out in stopping polio virus from their territory. Nigeria worked very efficiently and in a technical manner in achieving their goal, now they proved them self by eradicating the polio intelligently. All this can be done by the hard work of their government, their strategies and policies as well vaccinators and involvement of many brilliant minds behind it.

C. India

India eradicated the wild virus from its territory and claimed its eradication of polio around the globe. In India, they worked hard in disrupting the importation of wild polio virus from their region and sustained this interruption for 2 years to declaring its polio free region. They checked the infant and their adults and confirmed the negative reports in their all territories. They examined their water or wastage system on weekly basis and verified all the samples and got negative results. Hence, in 2012 "WHO" detached the India from endemic countries of the world.

When no chances of transmission occur in their region indicated the confirmation of polio virus. After complete examined the India after 2011, a certification of polio free countries is given. In the start, India used OPV vaccine instead of IPV. By using OPV, it will only weaken the polio virus germ but germs still remain in the body of effected one. Then they started the use of IPV, inactivated polio vaccine, they inactivated the wild polio germs in the body of patient and gradually all the germs becomes inactivated and destroyed. IPV did a remarkable achievement in the negative polio results. India most effected region was Bihar and Uttar Pradesh where mostly cases of wild polio virus in infant and adult was reported. Government of Bihar and Uttar Pradesh did their best in disrupting the polio virus. Finally, they achieved their target in disrupting the polio virus in their areas. During that time, they conducted more than ten times of vaccination campaign in a year to determine the status of conformation of polio virus and worked according to the outcomes of their coverage. Thus in their 2nd phase of eradication of polio virus cases, they scrutinized each and every infant of their respective regions and did their best in interrupting the wild polio virus.

The entire struggle behind the polio virus campaigns, they cleared their region with 0% of polio virus with IPV injected vaccine and hard work of day and night. True polio eradication required 0% chances of polio virus. India did this and proved their region completely polio free in the world [1, 10].

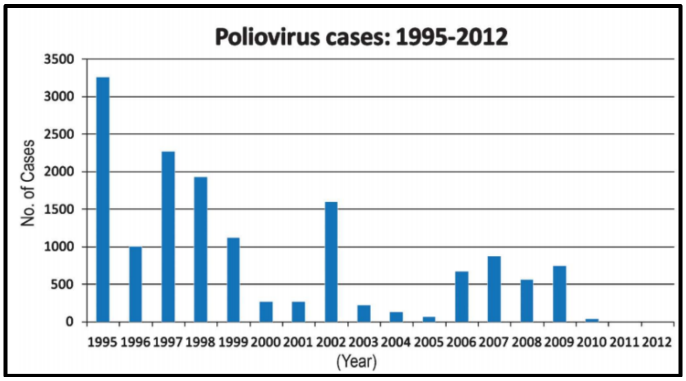


Fig. 2. Occurrence Wild polio cases in India

Finally, after the long struggling journey with polio India succeeded and declared as a polio free country. In last 3 years not even a single case reported from anywhere throughout the country regarding the wild poliovirus. Last case of poliovirus was reported from West Bengal State of India and the nature of the case was polio virus in January. Unfortunately polio virus has not been identified in sewage test in that span of time. India still needs to sustain his position for next 2 years to fully declare as polio free country. The main circumstances in India which makes the reason of polio cases had been the rejection of polio vaccine by few communities due to lack of education and propaganda against polio virus. On the other hand, regime of India playing a vital role by engaging/taking other stake holders on board which include media as well. Government of India also employed 2.5 million plus health workers who worked days and nights and had successfully achieved this landmark [11, 12].

III. SURVEYS

The primary goal of these surveys is to identify the problems in current polio vaccination system. The target populations of surveys are general public and vaccination teams. Another objective of these surveys is to find all those factors, policies, practices and reasons causing unsuccessful campaign and unsafe vaccination process and having polio virus in Pakistan.

We make sure that the survey queries are modified according to the specific groupings i.e. we ask non-technical questions from non-technical population like general public and similarly we ask technical questions from technical people like vaccinators or health care provider. These two types of target populations, we design two surveys with different questions in each questionnaire.

A. *General Public Survey*

In general public questionnaire, there are almost 51 questions which were asked from the parents of immunized child in last 6 months to 1 year. Hand written questionnaire was given to each person. This questionnaire is in-depth study of Polio vaccination system and its services. This report focus solely on the findings related to the vaccination system. Questionnaire results will serve to urge the presence of faults and lacking in the system where the further improvement can be made.

Survey Findings

After analyzed the survey data, we organized the data and highlights the important findings from general public survey, is as follows

- Polio is still the issue of Pakistan and its vaccination is necessary.
- Parents need more information to vaccinate their child.
- Parents missed the vaccination of their child mostly.
- Missed child could not be contacted for vaccination.
- Polio Vaccination system is older than current time and advancement of technology
- People spread wrong information and gives ill-advice to others.
- Polio Vaccination system is not safe and need to make the system safe for Polio workers.
- We have Polio in Pakistan because of insecurity.
- There should be more focus on danger zone areas.
- It is the need of time to make the system efficient and competent with respect to security and management

Parents don’t believe in the use of whether they use either from philosophical aspects reasons or as a consequence of religious concerns and beliefs, and similar proportion of such a people has concerns about the safety of usage of vaccines

B. *Vaccination Team Survey*

In vaccination team survey, there are almost 65 questions which were asked from the vaccinators and health workers. According to the survey, ten out of ten strongly believe that Polio is public health problem in Pakistan and all the children should be given Polio vaccine. Only three in ten attended/heard about Polio vaccination campaign. 100% vaccinators strongly believe that Vaccination eradication campaigns should be carried out. Seven out of ten strongly agree while 2 are agreeing on that there is still need to focus more on the activities towards Polio awareness campaign.

Survey Findings

After analyzed the survey data, we organized the data and highlights the important findings from general public survey, is as follows:

- Polio is still the issue of Pakistan and vaccination should be necessary and must be applicable in Pakistan.
- Vaccination eradication campaign needs to work more.
- There is still need to enhance the Polio awareness campaign.
- Vaccination system is older than current time and advancement of technology.
- Campaign need to be improved operationally and its management should be enhanced.
- Complaints of health workers are usually occur.
- No record of immunization found electronically
- Vaccination system does not have any strategy to tracked missed children.
- Vaccination has so many lacking on the subject of security, operations and communication
- Vaccination system is not secure and need to make the system safe for Polio workers.
- Danger zone areas need more attention and focus
- There is need to make the system efficient and competent with respect to the security and management
- Poor Security is the subject of matter of having Polio in Pakistan.
- Polio vaccination systems of Pakistan have not any proper technique to cover up those areas which are exposed to high risk.
- High risky areas are being neglected because of poor security and administration is unable to pay attention on danger zone areas.
- There is need to improve the way of campaign being conducted in risky areas
- Insecure vaccination system ruins the whole campaign and disheartens the workers.
- Security system requires improvement and enhancement.
- Poor security system is the reason of Polio cases in risky areas.
- People do not inject vaccine because of theoretical and religious mind set.

IV. FLAWS IDENTIFIED IN CURRENT PVS OF PAKISTAN

Followings flaws have been identified in current vaccination system after two full fledge surveys.

A. No/incomplete immunization record

Current polio vaccination system of Pakistan has a very poor record keeping process. Most of the time no record is maintained by the staff or fake entries is made. Even if staff maintains the record regularly, the data being kept is not enough to keep record of missed children or dual entries. Figure 3 and 4 are examples of current record keeping.

B. Paper based vaccination system

Vaccination records are completely maintained on papers. Child immunization registration is done on hand written paper. Paper based systems are prone to many human mistakes where there is no system of error feedback. Hence, the quality of data is uncertain and inaccurate. Figure 5 and 6 are examples of current record keeping.

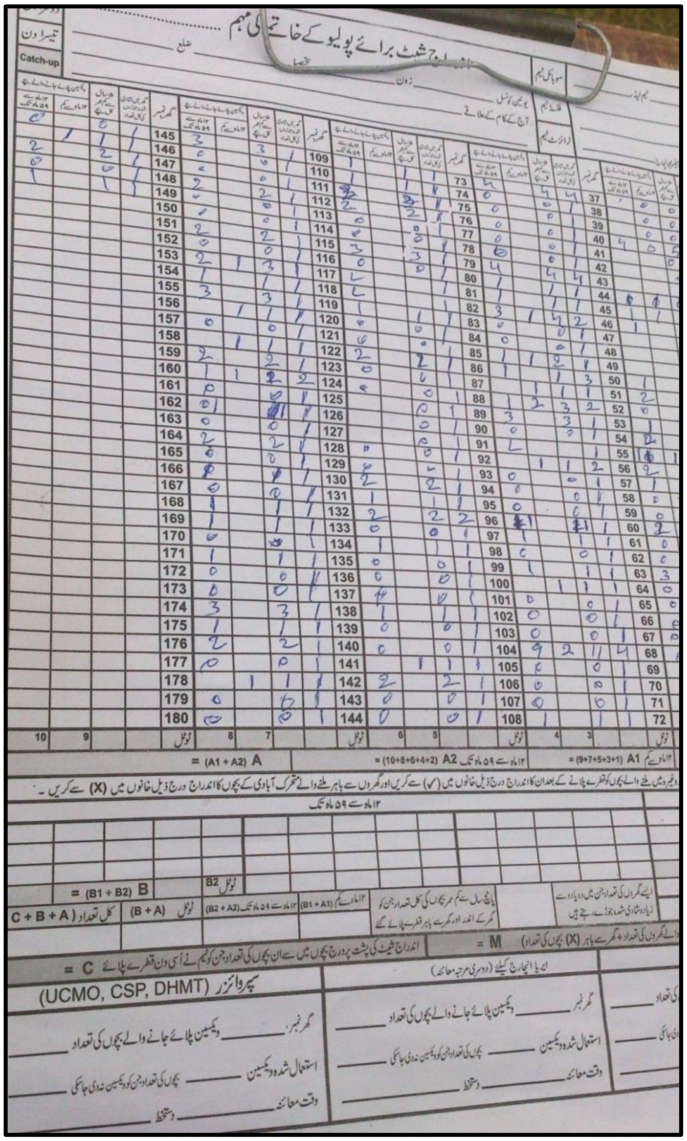

Catch-up
(UCMO, CSP, DHMT)

Fig. 3. Entry List for the Immunization Regarding Polio Eradication

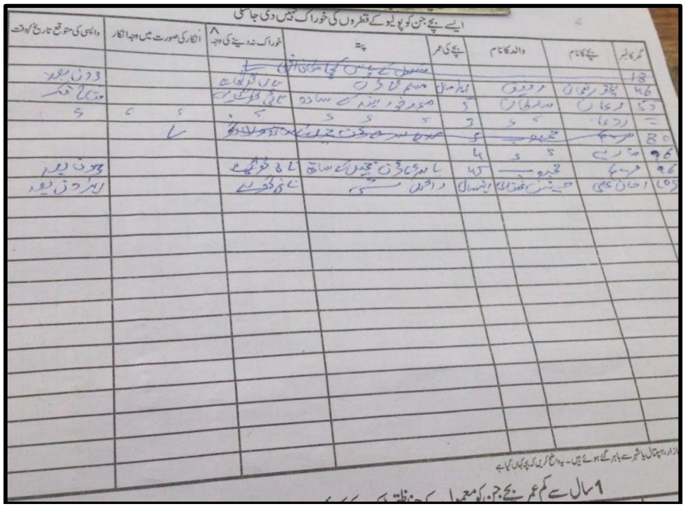


Fig. 4. Entry List for the Immunization Regarding Polio Eradication

C. High rate of missed children

During campaign, most of children remain deprived of vaccination or some are vaccinated twice. Poor record keeping and work corruption are main reasons of this problem. Unfortunately, there are no specific policies or laws being set which facilitate to immunize those missed children. This is the significant factor of rising polio cases in Pakistan.

D. Lack of Technology

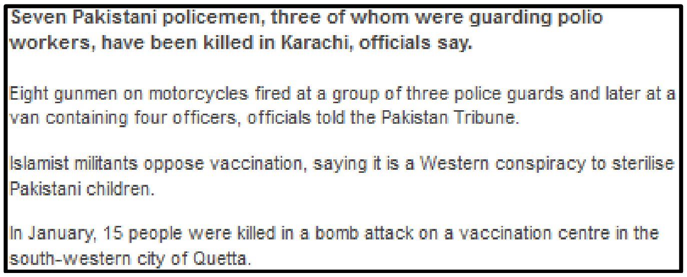
In current vaccination process, lack of technology limits the capacity and effectiveness of the health worker. There is speedily widespread adoption of new technology in immunization system in developing countries. It is the need of the time to push down the new technology in the system. Use of advanced technology in vaccination system, it will not only be the well-organized structure but also a proficient system.

E. Mistrustful information about Polio Campaigns

Misinformation has destabilized polio vaccine campaigns in Pakistan. Misconceptions and lack of education are considered as foremost barricades regarding polio vaccination. Parents are unaware of the threats of vaccine preventable disease. People especially in remote areas have so many misconceptions about vaccines. They fully negate to immunize, for not having a proper awareness about vaccine and its benefits. Thus, incredible information and misconception increase the barrier in the delivery of immunization.

F. No Security Zone in High Risk Areas

High security risk areas have denied access to some part of the population and hinder the growth of 'End Polio from Pakistan' campaign. This is one of the major reasons behind the presence of polio cases in Pakistan. In Pakistan the main challenges in scheming polio conduction consist of poor and inappropriate security situations and unreachable areas with geological barricade. Infants are usually unapproachable and health care takers face intricacy in upholding polio virus campaigns productively. Threatening and killing polio health workers discourage them in performing their duties. This leads to halt on the vaccination campaign. Such Mishaps dispirit the community and they refuse vaccination too. In 2012 Pakistani Taliban commander banned the polio vaccination in Federally Administrated Tribal Areas (FATA) of Pakistan in reaction of United States drone attacks. Most of the polio cases are also found in FATA, Khyber Pakhtunkhwa (KPK), Baluchistan, Karachi and its surrounding areas. In Baluchistan, numerous cases were reported in from Quetta division. This growing complexity in vaccination process increases the need of high security measures. Security risks continue in Karachi and its related areas. Thus it is very difficult to maintain the quality of campaign and to continue with it. Most recently case was occur in April, 2016 in Karachi an attacked on polio team during vaccination campaign, killed cops which were on duty as a guards of polio team.



Seven Pakistani policemen, three of whom were guarding polio workers, have been killed in Karachi, officials say.

Eight gunmen on motorcycles fired at a group of three police guards and later at a van containing four officers, officials told the Pakistan Tribune.

Islamist militants oppose vaccination, saying it is a Western conspiracy to sterilise Pakistani children.

In January, 15 people were killed in a bomb attack on a vaccination centre in the south-western city of Quetta.

Fig. 5. Attack on the polio team in April 2016 [13]

G. Poor performance of vaccinators

The role of health service providers in the eradication of polio virus program can't be overlooked. The vaccinators already know their assigned regions or community, the territory and the language of the area in which they work, facilitating the job of governing the vaccine with a high rate of coverage. They can better immunize the children and increase the rate of immunization. The efforts of millions of people from different communities and walks of life have made anti poliovirus activities successful by performing their duties with punctuality, dedication and sincerity. Unfortunately, most of the vaccinators cannot carry out their task properly. Leaving children without vaccination out of laziness or saving vaccines is one of the most reported problems. Thus, Proper corrective measures must be taken immediately to check the quality of immunization coverage, performance of vaccinators otherwise the eradication program may be adversely affected.

H. Inadequate management and Poor co-ordination

A good management in any system led to high beyond expectations. Current vaccination system in Pakistan has so much lacking, inadequate management from the health officers, medical staff and government is also one of the obstacles. Vaccination system has no appropriate and proper management. Poor co-ordination between vaccinators and senior staff ruins the quality of campaign. Inadequate supervision with the poor accountability of the vaccination campaign is the factors of the failure of immunization system. Unfortunately, polio reported cases, highlighting the deficiencies and negligence of health management system.

All these problems are causing polio still prevailing in Pakistan. There is a need for a technology-based system that can compensate most of these problems identified during surveys. In this paper, we propose such a polio vaccination information system which is based on mobile technology and helps overcoming all challenges.

The figure 6 depicts the process adopted during the existing vaccination system.

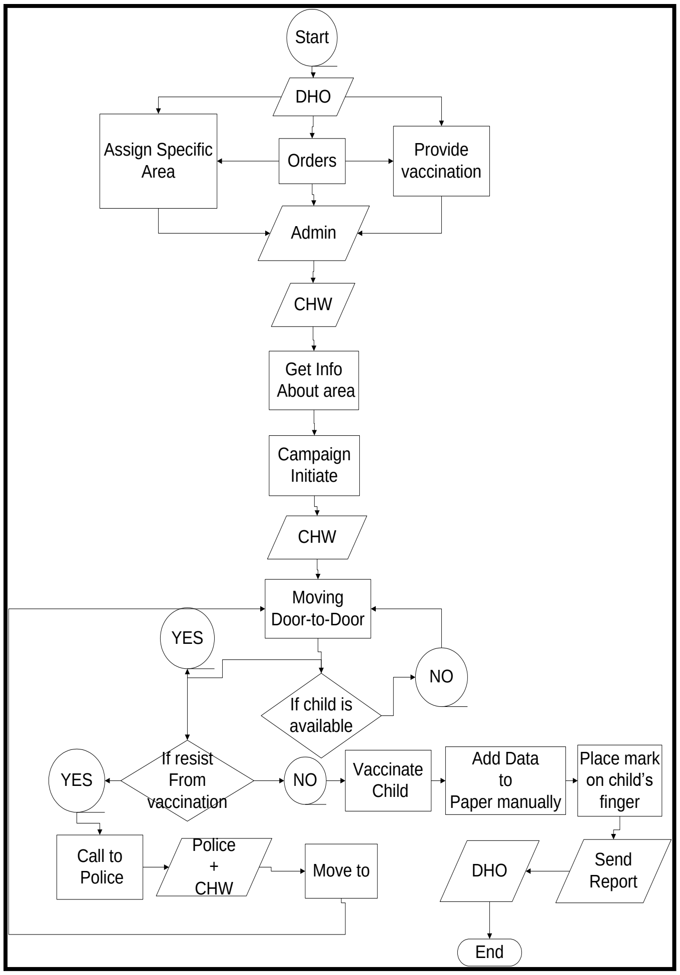


Fig. 6. Current Polio Vaccination System of Pakistan

In this model of current system, orders arrive from the District Health Quarter (DHQ) to the District Health Officer (DHO) about the campaign initiate. Then, orders along with schedule dates and vaccinations given to the administration of specific areas. Admin assign these to Community Health Workers (CHW) for the further vaccination process. CHW get information about their given areas then initiate campaign.

When campaign initiates, CHW move door-to-door, if child is available and at home then they are immunized by polio vaccine. Then CHW add trivial information of the vaccinated child on a sheet of paper (see figures 3 and 4) and put mark on child's finger. If someone shows resistance during immunization process like refusal of vaccine to their child then they simply inform the local police. This process of vaccination to the children will go on and health vaccinators visited house one by one in specific areas. During house-to-house visit, if child is not present at home then vaccinators visit again and again to check the availability of the child for the vaccination. CHW mentioned the absence of child on the given sheet of paper, during campaign interval. When campaign time period comes to an end then report submitted to the DHO.

Above model illustrate this whole scenario of current polio vaccination system. Campaign process is so simple and doesn't have any strategic tactics or advanced technology. There is not any sort of electronic data involvement. Sheet of paper is being used to write the temporary data for an instance. System doesn't have any new or advanced technique for the missed children. Top most is the unsafe working environment for the vaccinator. There is no any adequate security given to the health provider.

V. Proposed Technology Oriented PVS for Pakistan

This model represents the proposed polio vaccination system. Model represents the current vaccination system with the real time technology, where GPS, remainder recall system, tracking system and monitoring system involved. Proposed system involved electronic data, and data will be updated correspondingly.

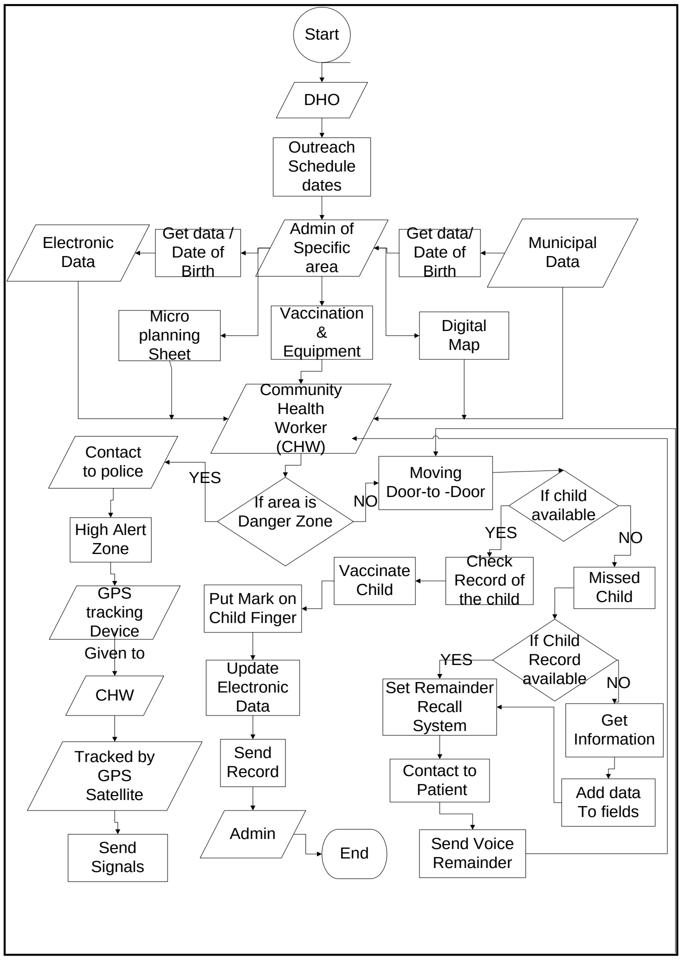


Fig. 7. (a) Proposed Scenario of Polio Vaccination System

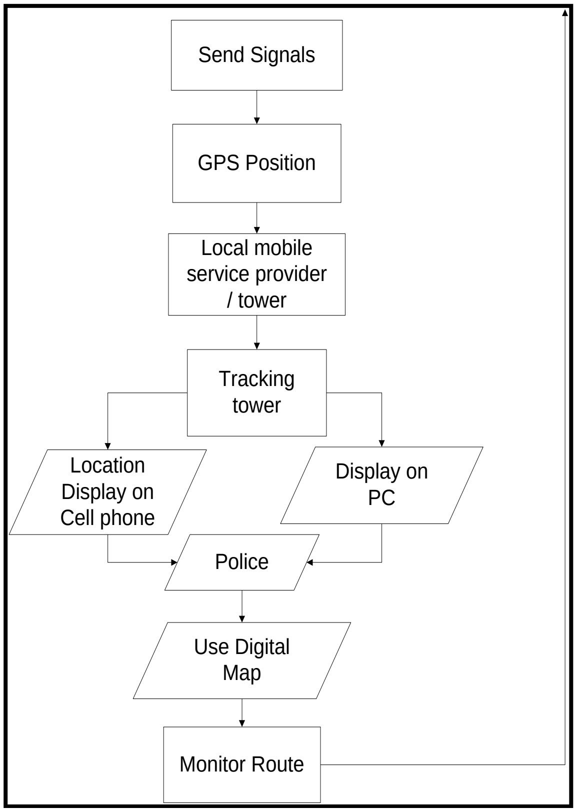


Fig. 7. (b) Proposed Scenario of Polio Vaccination System

The flow chart in figure 7-B represents the proposed polio vaccination system and its work flow. System starts with the schedule and orders from the DHO. Outreach program of immunization will initiate by administration according to the given dates. Administration of specific areas gets the data of infant from municipal. Now this data will be converted into electronic data for the further use and for the secure work. This electronic record along with its credentials given to the CHWs. Vaccination, micro planning and digital map will also be given t the CHWs. Micro planning sheets consist of detail work flow and digital map of their specific areas for their convenience as well as to cover up the whole area sequentially. Thus, CHW will move door-to door by having all these possessions. Most of the polio affected areas of Pakistan are high risky, such areas should be heighted as a danger zone and need high precautions of safety. So, it's better to check whether area is lying in danger zone region or not. If area lies out of the danger region, means area is safe there is no need high safety precaution. For the safe areas, CHWs move door-to-door, if child is available then CHWs will vaccinate the children and update the electronic record against the child's data. This updated record will be shown to the admin as well for the further monitoring. This process will be going on. If area is in danger zone then, CHWs will go with police for the immunization. High alert will be at local police quarter and administration will also keep checking status of vaccination through vaccinators and police department. A tracking system will be given to the vaccinators and through tracking system, vaccinators performance will be measure and will be shown to admin as well as local police head quarter. Signals of the tracking system will be shown to the local PC or cell phone. If any mishap will occur or vaccinators need more security, then through tracking signals police force with move with the digital map as well. So they reached on time and get over the worst situation easily. If child is not available at home, then child will be included in the list of missed children electronically. CHWs will check whether record against missing child is available or not. If record will be provided then CHWs sent voice remainder to their corresponding numbers. Re-contact the guardian or the patient again for the immunization. If no data will be present then, administration will get the information against the patient.

A. Re-Campaign Model of Proposed Polio vaccination system in Pakistan

In re-campaign polio vaccination system, re-campaign for the missed children are as important as other processes to be accomplished. High rate of missed children is one of the main reasons of having polio virus in Pakistan. Thus to get over it, an accurate and adequate re-campaign is the need of time. In re-campaign process, CHWs get the data of missed children from the electronic record. CHWs check the area during door-to-door vaccination process, whether more missed children is in remote areas or the city area. Then CHWs will move after getting the required information of the given area so they will perform better.

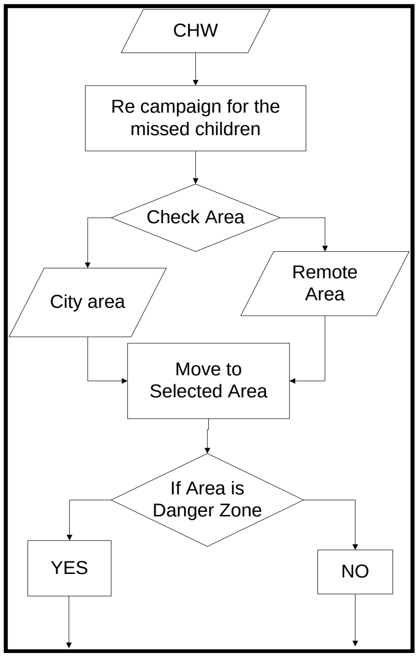


Fig. 8. Proposed Re-Campaign Scenario of Polio vaccination system

VI. Formal Modeling of Proposed PVS

A. Formal Verification

Formal verification efforts to give an answer for Functional check that sidesteps these issues by confirming (or refuting) accuracy of the configuration concerning the particular. A formal confirmation of a property is proportional to thorough testing of the configuration concerning that property [14].

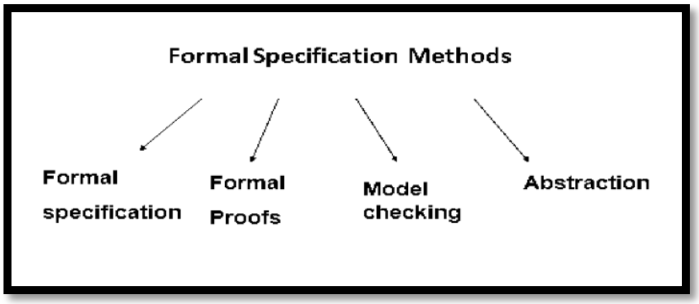


Fig. 9. Main Division of Formal Methods

Model Checking

Model checking has many kinds of displaying formalisms like Colored Petri Nets (CPNs) graphical demonstrating dialect. A CPN model is a dialect for displaying and acceptance of simultaneous and conveyed frameworks and different frameworks in which concurrency, synchronization, and correspondence assumes a noteworthy part [15, 16, 17].

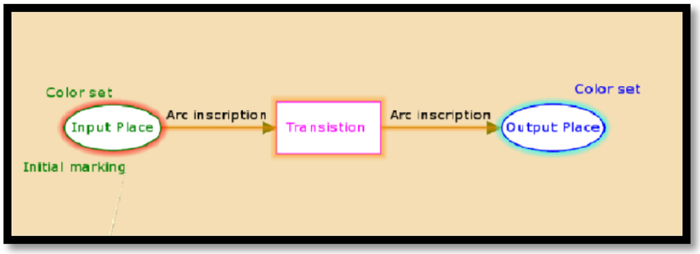


Fig. 10. Structure of Colored Petri Nets

Colored Petri Nets, where data is joined to every token. The data can be examined and changed when a move fires. For most applications, this speculation of customary Colored Petri Nets permits the client to make more sensible depictions, because of the way that equivalent sub Nets can be collapsed into one another, yielding a much littler Net. For reasons unknown spot invariants and achieve capacity trees, noteworthy techniques for standard CPNs, can be summed up to pertain for Colored Petri Nets [18].Colored Petri Nets and Predicate/transition-Nets are very intently linked, in the way that Colored Petri Nets have been established as an alteration of Predicate/transition-Nets, in order to avoid some practical troubles which rise when the technique of place-invariants is general to apply for Predicate [18].Colored Petri Net concept delivers potent analysis techniques used to verify the accuracy of workflow procedures. Acceptance of a model-driven approach, joined through comprehensive verification methods can make available a key solution for making qualitative and quantitative calculations about the possible system behaviors [18, 19].

Our work details a study in Verification of a system using Formal Modeling of Polio Vaccination System Using Colored Petri Nets. We show how support for features of correctness and verification in Colored Petri Nets enables demonstration [19, 20].

System verification is a central part in the software development process. It requires however, the system under the test of Colored Petri Nets to be at least moderately implemented. Also the practical verification of the systems working is exposed using simulations. Based on the usage of Colored Petri Nets as a specification tool we present an approach allowing the application of systems processes and transitions. As an additional benefit, the well-defined semantics of Colored Petri Nets enforces completeness and consistency of the system specification. The execution of the described technique is relied on the development of the vaccination system [21, 22]

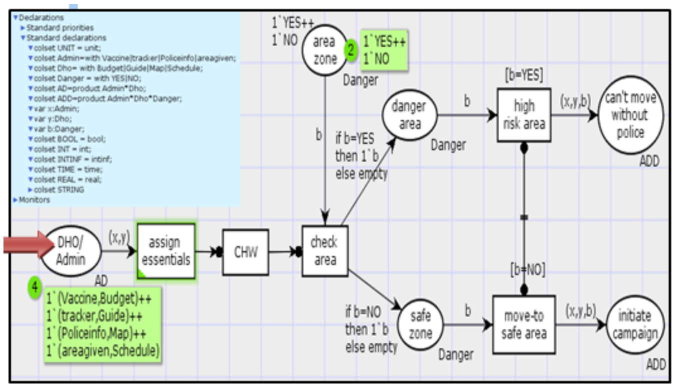
Functional Architecture

Colored Petri Nets models are well structured and provide planned modeling. Interactive simulations are created in using CPN tool. First we plot an idea to model the system using formal methods of software engineering [23, 24].

In Colored Petri Nets the conditions are depicted as a number of tokens presented as places and graphically represented in ellipse symbol. The data values are of different types like String, Integers, Unit, Boolean. We can combine some values with arithmetic operators. The actions are triggered in form of Transitions. In Colored Petri Nets that are designed the data flows through Arcs. The Arcs gets the Input from the tokens and process them in Transition as an Output. This model gives a complete strength on the conditions. For example, if we give a valid argument then we gets the desired result, otherwise the token will not flow. Green highlighted of transition shows that transition is active and flow work correctly [25, 26].

Area Zone Scenario

In figure 11-A, CPN of our system shows the basis scenario of area zone. DHO and Admin assign the basic essentials to the field worker to initiate the campaign. The new constrain used between CHW and check area is "new precedence constrain". This constrain will shows the precedence between transitions [27, 28]. Then field worker first check the area, whether health care provider move to safe area zone or danger region. In this colored Petri Net it is just start of simulation, when all tokens are ready to move by simulation with their prescribed variables and values [29, 30].



```
▼Declarations
  ▸Standard priorities
  ▼Standard declarations
    ▼colset UNIT = unit;
    ▼colset Admin=with Vaccine|tracker|Policeinfo|areagiven;
    ▼colset Dho= with Budget|Guide|Map|Schedule;
    ▼colset Danger = with YES|NO;
    ▼colset AD=product Admin*Dho;
    ▼colset ADD=product Admin*Dho*Danger;
    ▼var x:Admin;
    ▼var y:Dho;
    ▼var b:Danger;
    ▼colset BOOL = bool;
    ▼colset INT = int;
    ▼colset INTINF = intinf;
    ▼colset TIME = time;
    ▼colset REAL = real;
    ▸colset STRING
▸Monitors
```

Fig. 11. (a) Main Simulation for the Area Zone Scenario

In figure 11-B, we are illustrating the variable declaration of our Proposed Polio Vaccination System. Colset represents the declaration with it specific value and each colset have some variable. Each variable belongs to its corresponding colset.

In second process CHW will select one area, either safe or danger area, so we use "new not co-existence constraint". This constraint show that both transitions will not enable at the same time, only transition will exist at once. It shows that CHW will move either towards safe area or danger area.

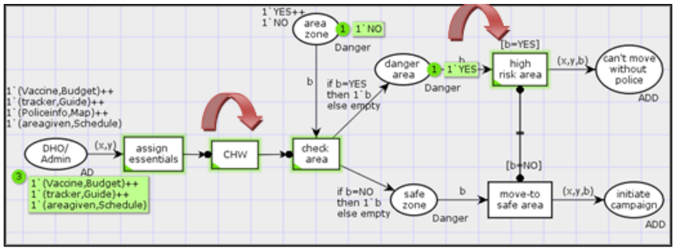


Fig. 11. (b) High risk area enables all transition

It shows that CHW will go to the danger area and only high risk area transition is enabled and all tokens passed and green highlighted the transition. Here value of variable "b= YES" depict that area is danger and CHW could not move without police. Output will be shown in the figure mentioned below.

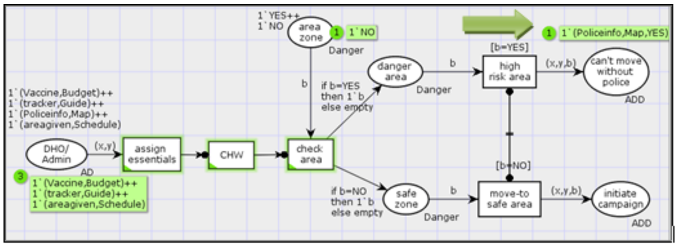


Fig. 11. (c) All token of High risk flow and shown in Output

In the above figure 11-C, output will be shown on the place. Where all token flows perfectly and depict that area is in danger zone and CHW can't move without police for the vaccination is shown with the green arrow.

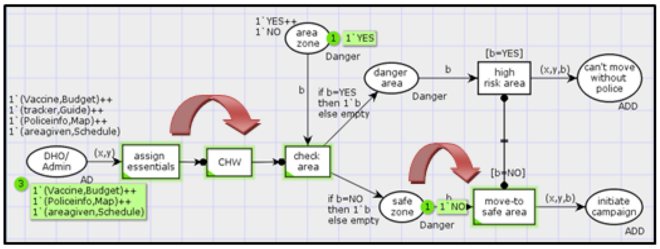


Fig. 11. (d) Safe area token followed and enabled all transition

In this CPN figure 11-D, it shows that CHW will go to the safe area and all the transition is enabled of the safer zone area, whereas "b=NO" represents that there is no danger, CHW can move in safe area zone and flow of tokens will be shown with red arrows in figure 11-E.

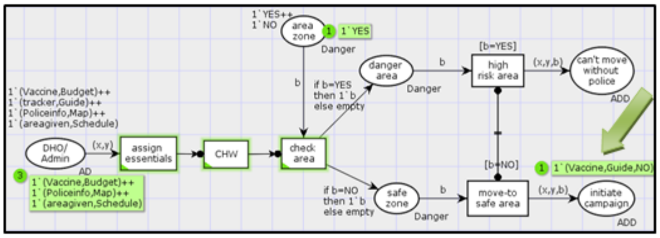
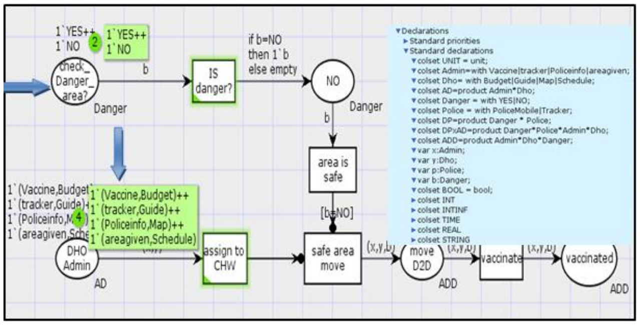


Fig. 11. (e) All tokens of Safe area flow and shown in Output

Safe Zone Scenario

In safe zone scenario of figure 12-A, this is the initial state of safe zone area, where no tokens are flowing. CHW will check if there is no danger then CHW will move, else tokens will not flow. Different variable are used in this scenario, "b" is the variable used to declare the danger, "x, y" variables are for the DHO and Admin, which will assign to the CHW to move door-to-door. In this Petri Net it is just start of simulation, when all tokens are ready to move by simulation with their prescribed variables and values.



```
▼Declarations
  ▶Standard priorities
  ▼Standard declarations
    ▼colset UNIT = unit;
    ▼colset Admin=with Vaccine|tracker|Policeinfo|areagiven;
    ▼colset Dho= with Budget|Guide|Map|Schedule;
    ▼colset AD=product Admin*Dho;
    ▼colset Danger = with YES|NO;
    ▼colset Police = with PoliceMobile|Tracker;
    ▼colset DP=product Danger * Police;
    ▼colset DPxAD=product Danger*Police*Admin*Dho;
    ▼colset ADD=product Admin*Dho*Danger;
    ▼var x:Admin;
    ▼var y:Dho;
    ▼var p:Police;
    ▼var b:Danger;
    ▼colset BOOL = bool;
    ▶colset INT
    ▶colset INTINF
    ▶colset TIME
    ▶colset REAL
    ▶colset STRING
```

Fig. 12. (a) Main Simulation for the Safe Area Zone Scenario

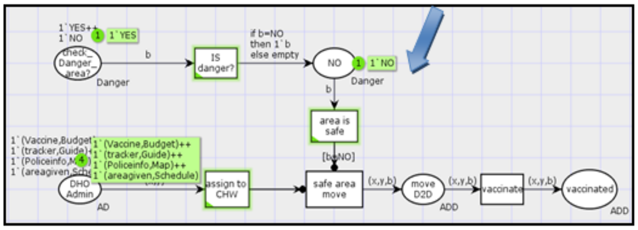


Fig. 12. (b) Simulation for No danger Zone Scenario

In above CPN figure 12-B, there is no danger the transition labeled with "area is safe" is highlighted and shows that token flowing accurately. If danger means "b=YES" then no token will flow and it will empty. "New Precedence constraint" is between transition shows the precedence level.

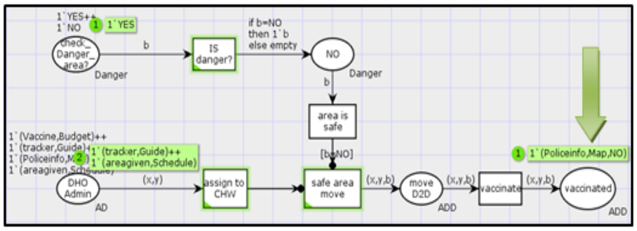
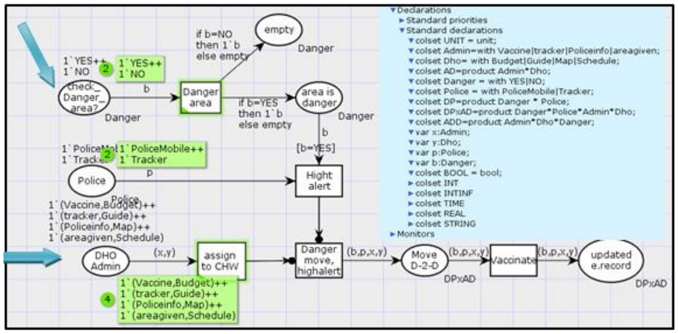


Fig. 12. (c) All tokens of Safe area flow and shown in Output

In this Figure 12-C, variable "b=NO", means safe area, so transition labeled with safe area is highlighted. All the tokens are flowing accurately and transition is enabled. Output will be shown in the last place.

Danger Zone Scenario

The first CPN figure 13-A represents the simulation of danger area and shows the basis scenario of danger area zone. There are more than one initial states of this CPN, first one is to check out that assign area is lies in danger zone or not. Second initial state is DHO and Admin assign the basic essentials to the field worker to initiate the campaign. The new constrain used around the transition labeled with "Danger move high alert" is "new precedence constrain". This constrain will shows the precedence between transitions. Then field worker first check the area, whether health care provider move to danger region and are high alert. In this Petri Net it is just start of simulation, when all tokens are ready to move by simulation with their prescribed variables and values.



```
▼Declarations
  ▶Standard priorities
  ▼Standard declarations
    ▼colset UNIT = unit;
    ▼colset Admin=with Vaccine|tracker|Policeinfo|areagiven;
    ▼colset Dho= with Budget|Guide|Map|Schedule;
    ▼colset AD=product Admin*Dho;
    ▼colset Danger = with YES|NO;
    ▼colset Police = with PoliceMobile|Tracker;
    ▼colset DP=product Danger * Police;
    ▼colset DPxAD=product Danger*Police*Admin*Dho;
    ▼colset ADD=product Admin*Dho*Danger;
    ▼var x:Admin;
    ▼var y:Dho;
    ▼var p:Police;
    ▼var b:Danger;
    ▼colset BOOL = bool;
    ▶colset INT
    ▶colset INTINF
    ▶colset TIME
    ▶colset REAL
    ▶colset STRING
  ▶Monitors
```

Fig. 13. (a) Main Simulation for the Danger Area Zone

This CPN model 13-A depicts the most of the functionality of our danger zone area, here are initial transitions represent danger area and CHW need police to go for the vaccination. It shows all tokens are in their initial states and ready to simulate. It shows all tokens are in their initial states and ready to simulate.

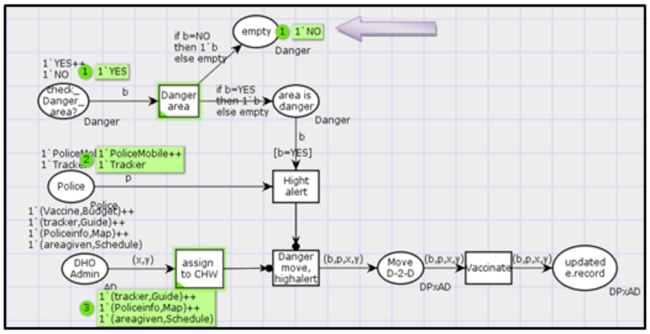


Fig. 13. (b) Simulation showing output with no danger in area

This CPN 13-B, initiate with the labeled "check danger area", CHW will check the area first to move for vaccination. This Figure 13-B, depict that if area is no danger, then it shows the empty place means no further transition or place going to be held. Tokens are flowing accurately and output shown on the place labeled with "empty".

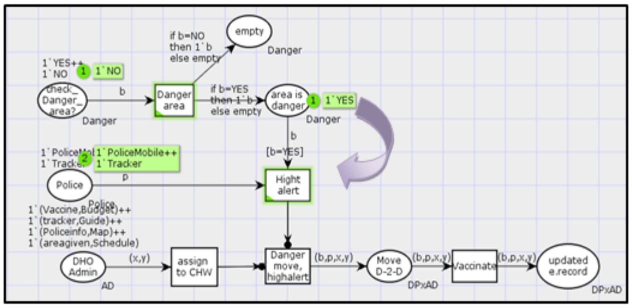


Fig. 13. (c) Simulation of High alert in danger area

If area will be in danger zone than high alert will green and enabled. In figure 13-C, shows danger area transition is shown highlighted means area is danger so police will interact with the health provider and high alert will be enabled.

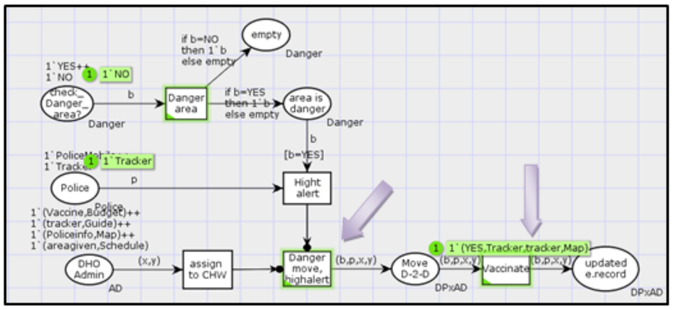


Fig. 13. (d) Simulation of move in danger area

Figure 13-D, shows all token flows perfectly and depict that area is in danger zone and high alerted. CHW move with police for the vaccination along with all essentials.

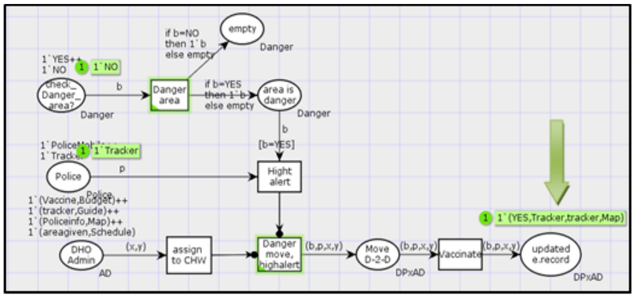


Fig. 13. (e) All token of Safe area flow and shown in Output

Final simulation of CPN 13-E is Showing the flow of all token from danger area if it will be "YES" and it highlights the "danger move, high alert" and token are functioning accurately and Petri Nets working perfectly. Green arrow shows the simulation of whole CPN module.

Child Availability Scenario

This CPN represents the scenario of availability of child. CHW will check the whether the child is available or not. If not available then child be in missed list, if child will be present at home for the vaccination then CHW will move further with the essentials. In this Petri Net it is just start of simulation, when all tokens are ready to move by simulation with their prescribed variables and values CPN shown with the red arrows in figures below.

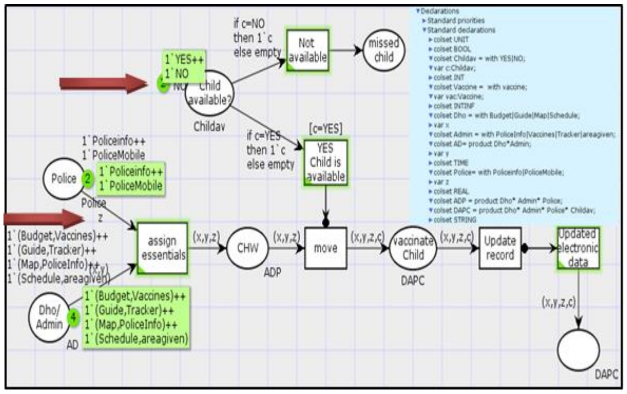


Fig. 14. (a) Main Simulation for the Child Availability Scenario

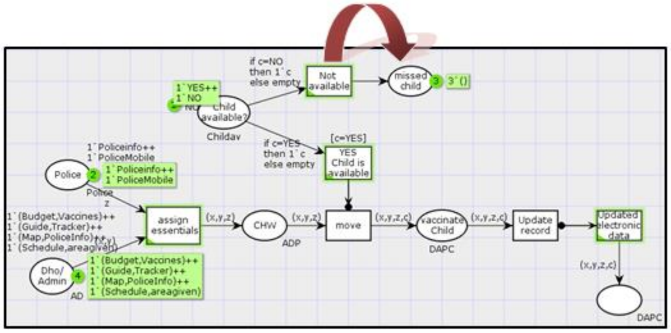


Fig. 14. (b) Simulation of missed child

In above CPN, it is shown that first tokens flowing to check the availability of child. Child is not available and transition is enabled labeled with "Not Available" and output shown in place of missed child.

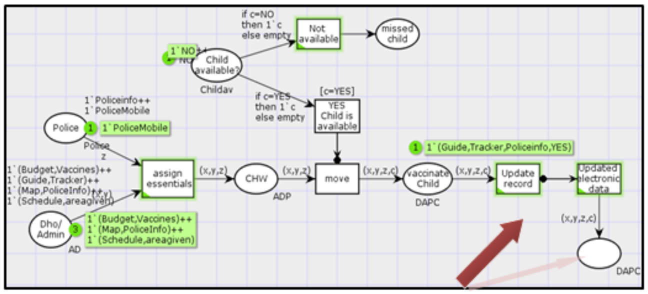


Fig. 14. (c) All token of Safe area flow and shown in Output

Figure 14-C shows simulation of vaccination child if available is exposed; transition is enabled shown with the green highlights. Token flowed accurately and Colored Petri Nets working correctly.

Missed Child Scenario

In this Petri Net it is just start of simulation, when all tokens are ready to move by simulation with their prescribed variables In this Missed Child CPN, CHW will check the whether the child is missed or not. If missed then child be in missed list, if child will not be present at home for the vaccination then CHW will move further and contact its relevant information for the vaccination and then update the record. In this scenario, there is new constrain between children available or not is “new not co-existence constraints” means transition both will not happened at the same time. If child will not be present then move to another transition and if not missed then move with different transition as shown in figures 15-A.

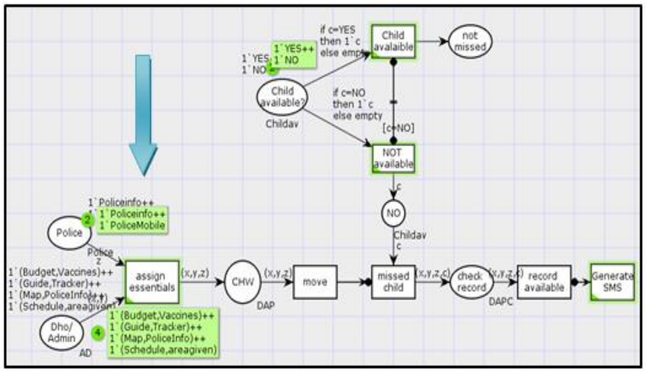


Fig. 15. (a) Main Simulation for the Missed Child Scenario

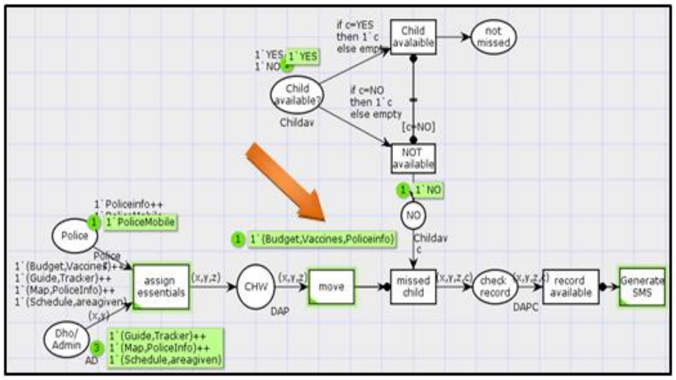


Fig. 15. (b) Simulation for the Missed Child with followed token

This figure 15-B shows that CHW will move with the essentials and check the presence of child first. If child is not available then “No” label and token flowed.

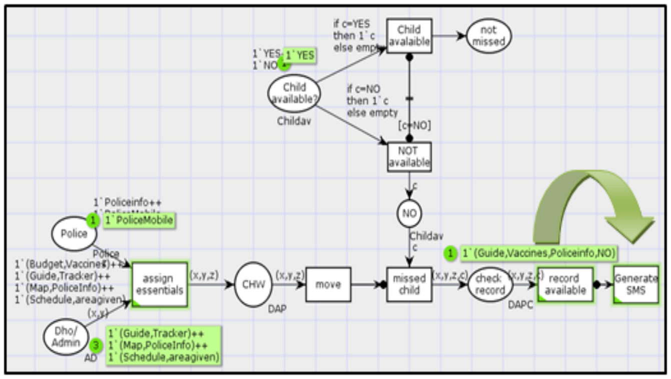


Fig. 15. (c) Simulation shown flow of token along with output

Figure 15-C shows simulation depicts the flow of tokens of missed child scenario and describes the most of the functionality of missed child picture. It shows all tokens are in their final states.

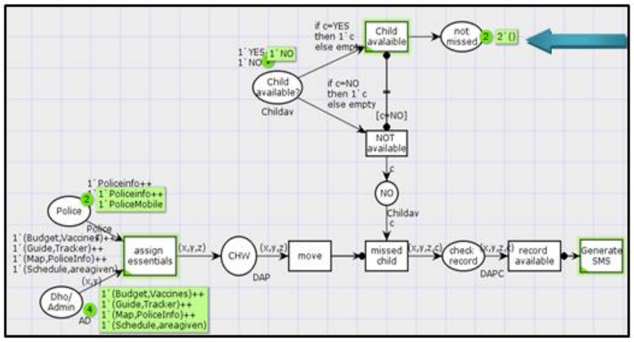


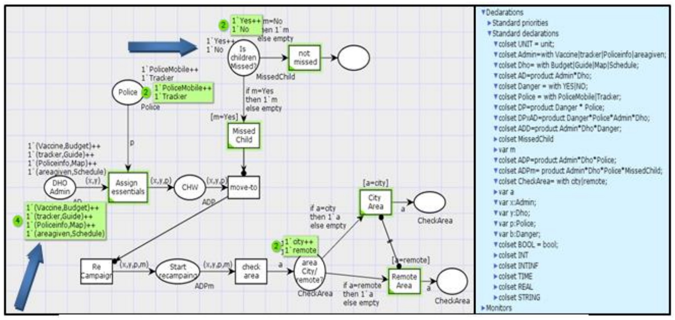
Fig. 15. (d) Simulation of the Not Missed Child

This CPN model 15-D, depicts that if child will be present at home then child will not be included in missed list, and transition is enabled and output is shown in the last output place.

Re-campaign Scenario

Colored Petri Nets based models are designed in many forms for research and formal modeling of different case studies Therefore, performance using Colored Petri Nets must rely on different simulations to show the complete performance measures for a supposed system.

In re-campaign CPN, CHW initiate the campaign with the missed children only. CHW initiate the campaign with all of its essentials and move to check the area if the given area is reside in city or it will be rural area. All tokens are ready to move by simulation with their prescribed variables and values, variables are initialized to accomplish the process for missed child and CPN shown in figure 16-A.



```
▾Declarations
  ▸Standard priorities
  ▾Standard declarations
    ▾colset UNIT = unit;
    ▾colset Admin=with Vaccine|tracker|Policeinfo|areagiven;
    ▾colset Dho= with Budget|Guide|Map|Schedule;
    ▾colset AD=product Admin*Dho;
    ▾colset Danger = with YES|NO;
    ▾colset Police = with PoliceMobile|Tracker;
    ▾colset DP=product Danger * Police;
    ▾colset DPxAD=product Danger*Police*Admin*Dho;
    ▾colset ADD=product Admin*Dho*Danger;
    ▸colset MissedChild
    ▸var m
    ▾colset ADP=product Admin*Dho*Police;
    ▾colset ADPm= product Admin*Dho*Police*MissedChild;
    ▾colset CheckArea= with city|remote;
    ▸var a
    ▾var x:Admin;
    ▾var y:Dho;
    ▾var p:Police;
    ▾var b:Danger;
    ▾colset BOOL = bool;
    ▸colset INT
    ▸colset INTINF
    ▸colset TIME
    ▸colset REAL
    ▸colset STRING
▸Monitors
```

Fig. 16. (a) Initial Simulation of the Not Missed Child

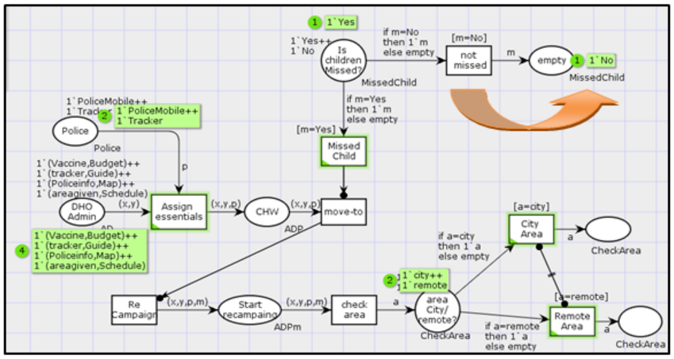


Fig. 16. (b) Simulation of the Not Missed Child

In this CPN 16-B, it shows the simulation of re-campaign. It shows whether child is missed or not, if not missed then it will be empty places. Tokens are flowing accurately and Colored Petri Nets are working correctly, shows with the pointed arrow.

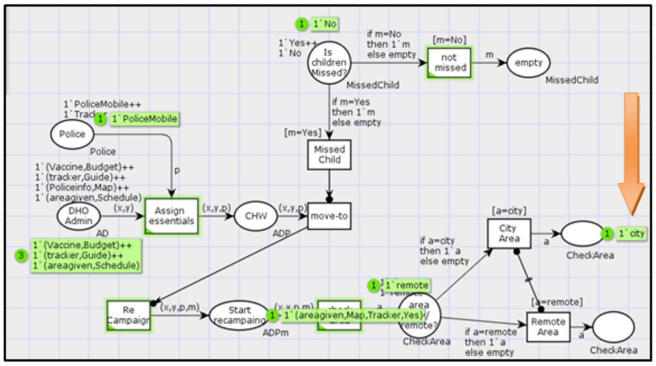


Fig. 16. (c) Simulation of re-campaign of city area

In this figure 16-C, shows the simulation of re-campaign of city area. This CPN model represents the flow of token and output represents with the sign of arrow.

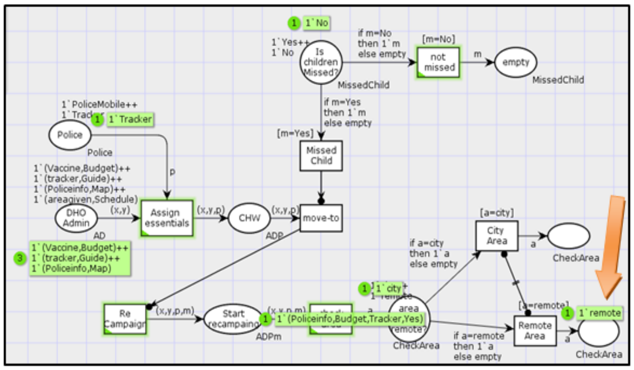


Fig. 16. (d) Simulation of re-campaign with remote area

Final simulation of this CPN 16-D, showing all the data is available in the form of output and remote area will be selected. This model is representing the flow of tokens with sign of arrow and transition is enabled.

VII. Conclusions

After scrutinized the existing polio vaccination system, we proposed technology oriented secure polio vaccination system. Our purpose is to build a system which makes the polio vaccination system effective and competent so we will be able to stand against the wild polio virus and make the Pakistan free from polio virus. The focus of this paper is directed towards the use of advanced tools in current vaccination system. The goal of the work offered is to commence a technology to make the polio vaccination system efficient in general and security perspective. This thesis investigate those tools which can be used to make a vaccination system effective and secure for the vaccinators which got death threat and find them defenseless in the fulfillment of their duties during vaccination campaign. Without proper security of health workers and general public it is not possible to carry out a successful campaign. On the other hand ourselves).by making Polio vaccination system's model we research a framework before we build it and to check the correctness of our proposed system by formal modeling. We also talked about the use of Colored Petri Nets to the Polio Vaccination system ends up being an application area which could profit by the elements of Colored Petri Nets. There are numerous great explanations behind utilizing Colored Petri Nets for data displaying and examination. CPN model is a depiction of displayed framework, and it could be utilized as an important feature (of a framework which we need to assemble) or as a presentation (of a framework which we need to disclose to other individuals. We did formal modeling of our Proposed technology oriented secure Polio Vaccination System. We verified our proposed model and by using Colored Petri Nets stimulate the different scenarios of our system. Hence we proved the correctness of our system.

References

[1] Kazi, Adbul Momin, M. Khalid, and A. N. Kazi. "Failure of polio eradication from Pakistan: Threat to world health." J Pioneer Med Sci 4.1 (2014): 8-9.

[2] GIS Mapping & GPS Tracking for Polio in Nigeria. Available from: http://mobile.thegatesnotes.com/Topics/Health/GIS-Mapping-GPS-Tracking-for-Polio-in-Nigeria

[3] Stockwell, Melissa S., and Alexander G. Fiks. "Utilizing health information technology to improve vaccine communication and coverage." Human vaccines & immunotherapeutics 9.8 (2013): 1802-1811.

[4] Shah, Syed Zawar, et al. "WHY WE COULD NOT ERADICATE POLIO FROM PAKISTAN AND HOW CAN WE?." Journal of Ayub Medical College Abbottabad 28.2 (2016): 423-425.

[5] Dutta, Anil. "Epidemiology of poliomyelitis—options and update." Vaccine 26.45 (2008): 5767-5773.

[6] Kay, Misha, Jonathan Santos, and Marina Takane. "mHealth: New horizons for health through mobile technologies." World Health Organization 64.7 (2011): 66-71.

[7] Bhaumik, Soumyadeep. "Polio eradication: Current status and challenges." Journal of family medicine and primary care 1.2 (2012): 84.

[8] Roush, Sandra W., Trudy V. Murphy, and Vaccine-Preventable Disease Table Working Group. "Historical comparisons of morbidity andmortality for vaccine-preventable diseases in the United States." Jama 298.18 (2007): 2155-2163.

[9] Barau, Inuwa, et al. "Improving polio vaccination coverage in Nigeria through the use of geographic information system technology." Journal of Infectious Diseases 210.suppl 1 (2014): S102-S110.

[10] Mehmood, Khalid, et al. "POLIO VACCINATION."

[11] Bhaumik, Soumyadeep. "Polio eradication: Current status and challenges." Journal of family medicine and primary care 1.2 (2012): 84.

[12] Kazi, A. M., et al. "Monitoring polio supplementary immunization activities using an automated short text messaging system in Karachi, Pakistan." Bulletin of the World Health Organization 92.3 (2014): 220-225.

[13] http://www.healthline.com/health/poliomyelitis.

[14] Parunak, H. Van Dyke. "Visualizing agent conversations: Using enhanced dooley graphs for agent design and analysis." Proceedings of the second international conference on multi-agent systems (ICMAS'96). 1996.

[15] Clarke, Edmund M., and E. Allen Emerson. "Design and synthesis of synchronization skeletons using branching time temporal logic." Workshop on Logic of Programs. Springer Berlin Heidelberg, 1981.

[16] Jensen, Kurt, and Lars M. Kristensen. Coloured Petri nets: modelling and validation of concurrent systems. Springer Science & Business Media, 2009.

[17] Brauer, Wilfried, Wolfgang Reisig, and Grzegorz Rozenberg, eds. Petri Nets: Central Models and Their Properties: Advances in Petri Nets 1986,

Part I Proceedings of an Advanced Course Bad Honnef, 8.–19. September 1986. Vol. 254. Springer, 2006.

[18] Emerson, E. Allen, and Jai Srinivasan. "Branching time temporal logic." Workshop/School/Symposium of the REX Project (Research and Education in Concurrent Systems). Springer Berlin Heidelberg, 1988.

[19] Nodine, Marian H., and Amy Unruh. "Facilitating open communication in agent systems: The infosleuth infrastructure." International Workshop on Agent Theories, Architectures, and Languages. Springer Berlin Heidelberg, 1997.

[20] Jensen, Kurt. Coloured Petri nets: basic concepts, analysis methods and practical use. Vol. 1. Springer Science & Business Media, 2013.

[21] L. Wells, S. Christensen, L. M. Kristensen, and K. Mortensen. Simulation based performance analysis of web servers. In R. German and B. Haverkort, editors, Proceedings of the 9th International Workshop on Petri Nets and Performance Models, pages 59–68. IEEE, 2001.

[22] B.Lindstrom and L. Wells. Annotating coloured Petri nets. To appear in the proceedings of the Fourth Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools (CPN'02), 2002.

[23] Christensen, S., Kristensen, L.M.: State Space Analysis of Hierarchical Coloured Petri Nets. In: Farwer, B., Moldt, D., Stehr, M.-O. (eds.): Proceedings of Workshop on Petri Nets in System Engineering { Modelling, Verification, and Validation. Department of Computer Science, University of Hamburg,1997, pp. 32{43, Report no. 20.

[24] Kristensen, Lars M., Soren Christensen, and Kurt Jensen. "The practitioner's guide to coloured Petri nets." International Journal on Software Tools for Technology Transfer (STTT) 2.2 (1998): 98-132.

[25] Gordon, Steven Donald. "Verification of the WAP transaction layer using coloured Petri nets." (2001).

[26] Jensen, Kurt. "Coloured Petri nets: A high level language for system design and analysis." International Conference on Application and Theory of Petri Nets. Springer Berlin Heidelberg, 1989.

[27] Zimmermann, Armin. "Colored petri nets." Stochastic Discrete Event Systems: Modeling, Evaluation, Applications (2008): 99-124.

[28] Cost, R. Scott, et al. "Modeling agent conversations with colored petri nets." Working Notes of the Workshop on Specifying and Implementing Conversation Policies. 1999.

[29] Liu, Dongsheng, et al. "Modeling workflow processes with colored Petri nets." computers in industry 49.3 (2002): 267-281.

[30] Cost, R. Scott, et al. "Using colored petri nets for conversation modeling." Issues in Agent Communication. Springer Berlin Heidelberg, 2000. 178-192.